IGAKU-SHOIN/SAUNDERS INTERNATIONAL EDITION

TEXTBOOK OF MEDICAL PHYSIOLOGY

SEVENTH EDITION

Arthur C. Guyton, M.D.
Chairman and Professor
Department of Physiology and Biophysics
University of Mississippi School of Medicine

1986
W. B. SAUNDERS COMPANY
IGAKU-SHOIN / SAUNDERS

Philadelphia London Toronto Mexico City
Rio de Janeiro Sydney Tokyo Hong Kong

W B Saunders Company: West Washington Square
Philadelphia, PA 19105

Library of Congress Cataloging-in-Publication Data

Guyton, Arthur C.
Textbook of medical physiology.
Includes bibliographies and index
1 Human physiology. I. Title. II. Title: Medical physiology. [DNLM. 1. Physiology. QT 104 G992t]
QP34.5.G9 1986 612 85-18294
ISBN 0-7216-1260-1

Listed here is the latest translated edition of this book together with the language of the translation and the publisher

Japanese (*4th Edition*)—Kirokawa Publishing Company, Tokyo, Japan
Serbo-Croatian (*4th Edition*)—Medicinska Knjiga, Belgrade, Yugoslavia
Italian (*4th Edition*)—Piccin Editore, Padova, Italy
Spanish (*6th Edition*)—Nueva Editorial Interamericana S A de C.V., Mexico City, Mexico
Persian (*5th Edition*)—Sherkat Sahami "TCHEHR," Tehran, Iran
Portuguese (*6th Edition*)—Editora Interamericana Ltda., Rio de Janeiro, Brazil

Editor Dana Dreibelbis
Designer Karen O'Keefe
Production Manager Bob Butler
Manuscript Editor Wynette Kommer
Illustrator Karen McGarry
Illustration Coordinator Peg Shaw

Dr. ... 1964 ... 12/11/57

Textbook of Medical Physiology

ISBN 0-7216-1260-1 (US Edition)
ISBN 4-7557-0113-9 (Igaku Shoin/ Saunders International Edition)

© 1986 by W B Saunders Company Copyright 1956, 1961, 1966, 1971, 1976 and 1981 by W B Saunders Company Copyright under the Uniform Copyright Convention Simultaneously published in Canada. All rights reserved This book is protected by copyright. No part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the publisher. Library of Congress catalog card number 85-18294.

This International Edition is not for sale in the United States of America, its dependencies or Canada

Printed in Japan, 1986

Last digit is the print number: 9 8 7 6 5 4 3 2 1

Dedicated to

MY FATHER
for the uncompromising principles that guided his life

MY MOTHER
for leading her children into intellectual pursuits

MY WIFE
for her magnificent devotion to her family

MY CHILDREN
for making everything worthwhile

Preface

When I wrote the first edition of this textbook, I had the naive belief that once the book was completed, subsequent revisions would require only simple changes. However, with each new edition I find that progress in the field of physiology is so rapid that large portions of the text must be completely recast and rewritten, and even the emphasis of the subject matter must often be changed as our knowledge becomes more penetrating. Therefore, the reader will find hardly a chapter in this edition that has not been significantly altered. Also, most of the figures have been either changed in at least some feature or replaced, and new ones have been added. Especially important, many of the figures have been redrawn simply to increase their clarity.

As has been true of most of the past revisions, the major change in this edition reflects a striking trend toward more fundamental physiology. The reason for this is mainly the greater success of the research physiologist in probing deeper into basic mechanisms of function, especially cellular and molecular mechanisms, than was true a few yeors ago. Yet I have attempted still to present physiology as an integrated study of the body's functional systems, utilizing the new fundamental knowledge to build a better understanding of the mechanisms upon which life depends.

Also a serious attempt has been made to devise and achieve tochniques of expression that will help bring tho physical and chemical principles of the body's complexities into the medical student's realm of understanding. In pursuit of this goal, I have kept records of those types of material with which the student has difficulty; I have quizzed students in detail to determine their levels of comprehension; and I have attempted to note inconsistencies in logic that might appear in student discussions. All these data have been used to help choose the material and methods of presentation. Thus, it has been my desire to make this book a "teaching" text as well as one that covers essentially all the basic physiology required of a student of medicine.

A special emphasis of the book is a more detailed attention than given in most other textbooks of physiology to the body's many control systems. These are the basis of what physiologists call *homeostasis*. The reason for this special emphasis is that most disease conditions of the body result from abnormal function of one or more of the control systems. Therefore, the student's comprehension of "medical" pbysiology depends perhaps more on a knowledge of these systems than on any other facet of physiology.

Another goal has been to make the text as accurate as possible. To help attain this, suggestions and critiques from many physiologists, students, and clinicians throughout the United States and other parts of the world have been received and utilized in checking the factual accuracy of the text. Yet, even so, because of the likelihood of error in sorting through many thousands of bits of information, I wish to issue an

invitation—in fact, much more than merely an invitation, actually a request—to all readers to send along notices of error or inaccuracy. Indeed, physiologists understand how important feedback is to proper function of the human body; so, too, is feedback equally important for progressive development of a textbook of medical physiology. I hope also that those many persons who have helped already will accept my sincerest appreciation for their efforts

A word of explanation is needed about two features of the text—first, the references, and, second, the two print sizes The references have been chosen primarily for their up-to-dateness and for the quality of their own bibliographies. Use of these references as well as of cross-references from them can give the student almost complete coverage of the entire field of physiology

The print is set in two sizes. The material presented in small print is of several different kinds; first, anatomical, chemical, and other information that is needed for the immediate discussion but that most students will learn in more detail in other courses, second, information that is of special importance to certain clinical fields of medicine but that is not necessary to fundamental understanding of the body's basic physiologic mechanisms; and, third, information that will be of value to those students who wish to pursue a subject more deeply than the average medical student In contrast, the material in large print constitutes the major bulk of physiological information that students will require in their medical studies and that they will not obtain in other courses. Those teachers who would like to present a limited course of physiology can direct students' study primarily to the large type

Again, I wish to express my deepest appreciation to many others who have helped in the preparation of this book. I am particularly grateful to Ivadelle Osberg Heidke, Gwendolyn Robbins, and Elaine Steed-Davis for their excellent secretarial services, to Tomiko Mita for her superb work on new illustrations, and to Dana Dreibelbis, Wynette Kommer, Robert Butler, Margaret Shaw, and the staff of the W. B. Saunders Company for continued editorial and production excellence.

ARTHUR C. GUYTON

Contents

I INTRODUCTION TO PHYSIOLOGY: THE CELL AND GENERAL PHYSIOLOGY 1

1 FUNCTIONAL ORGANIZATION OF THE HUMAN BODY AND CONTROL OF THE "INTERNAL ENVIRONMENT" 2
Cells as the Living Units of the Body 2
The Extracellular Fluid—The Internal Environment 2
"Homeostatic" Mechanisms of the Major Functional Systems 3
Homeostasis 3
The Extracellular Fluid Transport System—The Circulatory System 3
Origin of Nutrients in the Extracellular Fluid 4
Removal of Metabolic End-Products 4
Regulation of Body Functions 4
Reproduction 5
The Control Systems of the Body 5
Examples of Control Mechanisms 5
Characteristics of Control Systems 5
Automaticity of the Body 7
Appendix 7
Basic Physical Principles of Control Systems 7

2 THE CELL AND ITS FUNCTION 11
Organization of the Cell 11
Physical Structure of the Cell 12
The Membranous Structures of the Cell 12
The *Cytoplasm* and *Its Organelles* 14
Comparison of the Animal Cell with Precellular Forms of Life 18
Functional Systems of the Cell 19
Ingestion by the Cell—Endocytosis 19
Digestion of Foreign Substances in the Cell—Function of the Lysosomes 20
Synthesis and Formation of Cellular Structures by the Endoplasmic Reticulum and the Golgi Apparatus 21
Extraction of Energy from Nutrients—Function of the Mitochondria 22
Ameboid Locomotion of Cells 24
Cilia and Ciliary Movements 25

3 GENETIC CONTROL OF PROTEIN SYNTHESIS, CELL FUNCTION, AND CELL REPRODUCTION 27
The Genes 27
The Genetic Code 28
Ribonucleic Acid (RNA)—The Process of Transcription 29
Messenger RNA—The "Codons" 30
Transfer RNA—The "Anticodons" 31
Ribsomal RNA 31
Formation of Proteins on the Ribosomes—The Process of Translation 31
Synthesis of Other Substances in the Cell 33
Control of Genetic Function and Biochemical Activity in Cells 33
Genetic Regulation 33
Control of Enzyme Activity 34
Cell Reproduction 35
Replication of the DNA 35
The Chromosomes and Their Replication 36
Mitosis 36
Control of Cell Growth and Reproduction 37
Cell Differentiation 37
Cancer 38

II BLOOD CELLS, IMMUNITY, AND BLOOD CLOTTING 41

1 RED BLOOD CELLS, ANEMIA, AND POLYCYTHEMIA 42
The Red Blood Cells 42
Production of Red Blood Cells 42
Formation of Hemoglobin 46
Iron Metabolism .....
Destruction of Red Blood Cells .....
The Anemias .....

Effects of Anemia on the Circulatory System 49
Polycythemia 49
Effect of Polycythemia on the Circulatory System 49

5 RESISTANCE OF THE BODY TO INFECTION—THE LEUKOCYTES, THE MACROPHAGE SYSTEM, AND INFLAMMATION 51
The Leukocytes (White Blood Cells) 51
General Characteristics of Leukocytes 51
Genesis of the Leukocytes 52
Life Span of the White Blood Cells 52
Properties of Neutrophils, Monocytes, and Macrophages 53
The Tissue Macrophage System (The Reticuloendothelial System) 54
Inflammation and Function of Neutrophils and Macrophages 56
The Process of Inflammation 56
Neutrophilia Caused by Conditions Other Than Inflammation 57
The Eosinophils 57
The Basophils 58
Agranulocytosis 58
The Leukemias 58
Effects of Leukemia on the Body 59

6 IMMUNITY AND ALLERGY 60
Innate Immunity 60
Acquired Immunity 60
Two Basic Types of Acquired Immunity 60
Antigens 60
Role of Lymphoid Tissue in Acquired Immunity 61
Preprocessing of the T and B Lymphocytes 61
Specificity of Antibodies and T Lymphocytes—Role of Lymphocyte Clones 62
Origin of the Many Clones of Lymphocytes 62
Specific Attributes of the B Lymphocyte System—Humoral Immunity and the Antibodies 63
Special Attributes of the T Lymphocyte System-Activated Cells and "Cell-Mediated Immunity" 65
Vaccination 67
Passive Immunity 67
Allergy 67
An Allergy That Occurs in Normal People Delayed-Reaction Allergy 68
Allergies in the "Allergic" Person 68

7 BLOOD GROUPS; TRANSFUSION; TISSUE AND ORGAN TRANSPLANTATION 70
Antigenicity and Immune Reactions of Blood 70
O-A-B Blood Groups 70
The A and B Antigens—Called "Agglutinogens" 70
The Agglutinins 71
The Agglutination Process in Transfusion Reactions 71
Blood Typing 72
The Rh Blood Types 72
The Rh Immune Response 72
Other Blood Factors 73
Transfusion 73
Transfusion Reactions Resulting from Mismatched Blood Groups 73
Transplantation of Tissues and Organs 74
Attempts to Overcome the Antigen-Antibody Reactions in Transplanted Tissue 74

8 HEMOSTASIS AND BLOOD COAGULATION 76
Events in Hemostasis 76
Vascular Spasm 76
Formation of the Platelet Plug 76
Blood Coagulation in the Ruptured Vessel 77
Fibrous Organization or Dissolution of the Blood Clot 77
Mechanism of Blood Coagulation 77
Conversion of Prothrombin to Thrombin 78
Conversion of Fibrinogen to Fibrin—Formation of the Clot 78
The Vicious Circle of Clot Formation 79
Block of Clot Growth by Blood Flow 79
Initiation of Coagulation: Formation of Prothrombin Activator 79
Prevention of Blood Clotting in the Normal Vascular System—The Intravascular Anticoagulants 82
Lysis of Blood Clots—Plasmin 82
Conditions That Cause Excessive Bleeding in Human Beings 83
Decreased Prothrombin, Factor VII, Factor IX, and Factor X Caused by Vitamin K Deficiency 83
Hemophilia 83
Thrombocytopenia 83
Thromboembolic Conditions in the Human Being 84
Femoral Thrombosis and Massive Pulmonary Embolism 84
Disseminated Intravascular Clotting 84
Anticoagulants for Clinical Use 85
Heparin as an Intravenous Anticoagulant 85
Coumarins as Anticoagulants 85
Prevention of Blood Coagulation Outside the Body 85
Blood Coagulation Tests 85
Bleeding Time 85
Clotting Time 85
Prothrombin Time 85

III MEMBRANE PHYSIOLOGY, NERVE, AND MUSCLE 87

9 TRANSPORT THROUGH THE CELL MEMBRANE 88
Diffusion 89
Diffusion Through the Cell Membrane. 89
Net Diffusion Through the Protein Channels of the Cell Membrane and Factors That Affect It 93
Osmosis Across Selectively Permeable Membranes—Net Diffusion of Water... 94
Active Transport 97
Basic Mechanism of Active Transport.. 97
Secondary Active Transport: Sodium "Cotransport" of Glucose and Amino Acids 99
Active Transport Through Cellular Sheets 99

10 MEMBRANE POTENTIALS AND ACTION POTENTIALS 101
Basic Physics of Membrane Potentials . .. 101
Membrane Potentials Caused by Diffusion 101
Membrane Potentials Caused by Active Transport—The Sodium-Potassium "Electrogenic Pump". 103
Measuring the Membrane Potential .. 103
The Cell Membrane as an Electrical Capacitor 104
The Resting Membrane Potential of Nerves 104
Origin of the Normal Resting Membrane Potential 105
The Nerve Action Potential 106
The Voltage-Gated Sodium and Potassium Channels 106
Summary of the Events That Cause the Action-Potential 108
Roles of Other Ions During the Action Potential 109
Initiation of the Action Potential.... . 110
Propagation of the Action Potential 110
"Recharging" the Fiber Membrane-After Action Potentials—Importance of Energy Metabolism 111
The Spike Potential and the After-Potentials 112
Plateau in Some Action Potentials..... 113
Rhythmicity of Certain Excitable Tissues—Repetitive Discharge 113
Special Aspects of Signal Transmission in Nerve Trunks 114
Velocity of Conduction in Nerve Fibers 115
Excitation—The Process of Eliciting the Action Potential 115
Inhibition of Excitability—"Stabilizers" and Local Anesthetics... 117
Recording Membrane Potentials and Action Potentials 117

11 CONTRACTION OF SKELETAL MUSCLE 120
Physiologic Anatomy of Skeletal Muscle .. 120
The Skeletal Muscle Fiber 120
Molecular Mechanism of Muscle Contraction 122
Molecular Characteristics of the Contractile Filaments 123
Degree of Actin and Myosin Filament Overlap—Effect on Tension Developed by the Contracting Muscle 125
Relation of Velocity of Contraction to Load 126
Initiation of Muscle Contraction: Excitation-Contraction Coupling 126
The Muscle Action Potential 127
Release of Calcium Ions by the Sarcoplasmic Reticulum 127
The Source of Energy for Muscle Contraction 129
Characteristics of a Single Muscle Twitch 130
Mechanics of Skeletal Muscle Contraction 131
The Motor Unit 131
Summation of Muscle Contraction ... 131
Skeletal Muscle Tone 132
Muscle Fatigue 133
The Lever Systems of the Body 133
Special Features and Abnormalities of Skeletal Muscle Function 133
Muscle Hypertrophy 133
Muscle Atrophy 134
Prevention of Muscle Atrophy by Electrical Stimulation 134
Physical Contracture of Muscle Following Denervation 134
Rigor Mortis 134
Familial Periodic Paralysis 134
The Electromyogram 134

12 NEUROMUSCULAR TRANSMISSION; FUNCTION OF SMOOTH MUSCLE 136
Transmission of Impulses from Nerves to Skeletal Muscle Fibers: The Neuromuscular Junction 136
Myasthenia Gravis 139
Contraction of Smooth Muscle 139
Types of Smooth Muscle 139
The Contractile Process in Smooth Muscle 140
Membrane Potentials and Action Potentials in Smooth Muscle 141
Excitation-Contraction Coupling—Role of Calcium Ions 143
Neuromuscular Junctions of Smooth Muscle 144
Smooth Muscle Contraction Without Action Potentials—Effect of Local Tissue Factors and Hormones 145
Mechanical Characteristics of Smooth Muscle Contraction 146

Effects of Anemia on the Circulatory System 49
Polycythemia 49
Effect of Polycythemia on the Circulatory System 49

5 RESISTANCE OF THE BODY TO INFECTION—THE LEUKOCYTES, THE MACROPHAGE SYSTEM, AND INFLAMMATION 51
The Leukocytes (White Blood Cells) 51
General Characteristics of Leukocytes 51
Genesis of the Leukocytes 52
Life Span of the White Blood Cells 52
Properties of Neutrophils, Monocytes, and Macrophages 53
The Tissue Macrophage System (The Reticuloendothelial System) 54
Inflammation and Function of Neutrophils and Macrophages 56
The Process of Inflammation 56
Neutrophilia Caused by Conditions Other Than Inflammation 57
The Eosinophils 57
The Basophils 58
Agranulocytosis 58
The Leukemias 58
Effects of Leukemia on the Body 59

6 IMMUNITY AND ALLERGY 60
Innate Immunity 60
Acquired Immunity 60
Two Basic Types of Acquired Immunity 60
Antigens 60
Role of Lymphoid Tissue in Acquired Immunity 61
Preprocessing of the T and B Lymphocytes 61
Specificity of Antibodies and T Lymphocytes—Role of Lymphocyte Clones 62
Origin of the Many Clones of Lymphocytes 62
Specific Attributes of the B Lymphocyte System—Humoral Immunity and the Antibodies 63
Special Attributes of the T Lymphocyte System-Activated Cells and "Cell-Mediated Immunity" 65
Vaccination 67
Passive Immunity 67
Allergy 67
An Allergy That Occurs in Normal People: Delayed-Reaction Allergy 68
Allergies in the "Allergic" Person 68

7 BLOOD GROUPS, TRANSFUSION; TISSUE AND ORGAN TRANSPLANTATION 70
Antigenicity and Immune Reactions of Blood 70
O-A-B Blood Groups 70
The A and B Antigens—Called "Agglutinogens" 70
The Agglutinins 71
The Agglutination Process in Transfusion Reactions 71
Blood Typing 72
The Rh Blood Types 72
The Rh Immune Response 72
Other Blood Factors 73
Transfusion 73
Transfusion Reactions Resulting from Mismatched Blood Groups 73
Transplantation of Tissues and Organs 74
Attempts to Overcome the Antigen-Antibody Reactions in Transplanted Tissue 74

8 HEMOSTASIS AND BLOOD COAGULATION 76
Events in Hemostasis 76
Vascular Spasm 76
Formation of the Platelet Plug 76
Blood Coagulation in the Ruptured Vessel 77
Fibrous Organization or Dissolution of the Blood Clot 77
Mechanism of Blood Coagulation 77
Conversion of Prothrombin to Thrombin 78
Conversion of Fibrinogen to Fibrin—Formation of the Clot 78
The Vicious Circle of Clot Formation 79
Block of Clot Growth by Blood Flow 79
Initiation of Coagulation: Formation of Prothrombin Activator 79
Prevention of Blood Clotting in the Normal Vascular System—The Intravascular Anticoagulants 82
Lysis of Blood Clots—Plasmin 82
Conditions That Cause Excessive Bleeding in Human Beings 83
Decreased Prothrombin, Factor VII, Factor IX, and Factor X Caused by Vitamin K Deficiency 83
Hemophilia 83
Thrombocytopenia 83
Thromboembolic Conditions in the Human Being 84
Femoral Thrombosis and Massive Pulmonary Embolism 84
Disseminated Intravascular Clotting 84
Anticoagulants for Clinical Use ..... 8[illegible]
Heparin as an Intravenous Anticoagulant ..... 8[illegible]
Coumarins as Anticoagulants ..... 8[illegible]
Prevention of Blood Coagulation Outside the Body ..... 8[illegible]
Blood Coagulation Tests .....
Bleeding Time .....
Clotting Time .....
Prothrombin Time .....

III MEMBRANE PHYSIOLOGY, NERVE, AND MUSCLE 87

9 TRANSPORT THROUGH THE CELL MEMBRANE 88
Diffusion 89
Diffusion Through the Cell Membrane. 89
Net Diffusion Through the Protein Channels of the Cell Membrane and Factors That Affect It 93
Osmosis Across Selectively Permeable Membranes—Net Diffusion of Water... 94
Active Transport 97
Basic Mechanism of Active Transport.. 97
Secondary Active Transport: Sodium "Cotransport" of Glucose and Amino Acids 99
Active Transport Through Cellular Sheets 99

10 MEMBRANE POTENTIALS AND ACTION POTENTIALS 101
Basic Physics of Membrane Potentials 101
Membrane Potentials Caused by Diffusion 101
Membrane Potentials Caused by Active Transport—The Sodium-Potessium "Electrogenic Pump" 103
Measuring the Membrane Potential .. 103
The Cell Membrane as en Electrical Capacitor 104
The Resting Membrene Potential of Nerves 104
Origin of the Normal Resting Membrane Potential 105
The Nerve Action Potential 106
The Voltage-Gated Sodium and Potessium Channels 106
Summary of the Events That Cause the Action Potential 108
Roles of Other Ions During the Action Potential 109
Initiation of the Action Potential..... 110
Propagation of the Action Potential 110
"Recharging" the Fiber Membrane-After Action Potentials—Importance of Energy Metabolism 111
The Spike Potential and the After-Potentials 112
Plateau in Some Action Potentials..... 113
Rhythmicity of Certain Excitable Tissues—Repetitive Discharge..... 113
Special Aspects of Signal Transmission in Nerve Trunks 114
Velocity of Conduction in Nerve Fibers 115
Excitation—The Process of Eliciting the Action Potential 115
Inhibition of Excitability—"Stabilizers" and Local Anesthetics... 117
Recording Membrane Potentials and Action Potentials 117

11 CONTRACTION OF SKELETAL MUSCLE 120
Physiologic Anatomy of Skeletal Muscle .. 120
The Skeletal Muscle Fiber 120
Molecular Mechanism of Muscle Contraction 122
Molecular Characteristics of the Contractile Filaments 123
Degree of Actin and Myosin Filament Overlap—Effect on Tension Developed by the Contracting Muscle 125
Relation of Velocity of Contraction to Load 126
Initiation of Muscle Contraction: Excitation-Contraction Coupling 126
The Muscle Action Potential..... 127
Release of Calcium Ions by the Sarcoplasmic Reticulum 127
The Source of Energy for Muscle Contraction 129
Characteristics of a Single Muscle Twitch 130
Mechanics of Skeletal Muscle Contraction 131
The Motor Unit 131
Summation of Muscle Contraction 131
Skeletel Muscle Tone 132
Muscle Fatigue 133
The Lever Systems of the Body..... 133
Special Features and Abnormelities of Skeletal Muscle Function 133
Muscle Hypertrophy 133
Muscle Atrophy 134
Prevention of Muscle Atrophy by Electrical Stimulation 134
Physical Contracture of Muscle Following Denervation 134
Rigor Mortis 134
Familial Periodic Paralysis 134
The Electromyogram 134

12 NEUROMUSCULAR TRANSMISSION; FUNCTION OF SMOOTH MUSCLE 136
Transmission of Impulses from Nerves to Skeletal Muscle Fibers: The Neuromuscular Junction 136
Myasthenia Gravis 139
Contraction of Smooth Muscle 139
Types of Smooth Muscle 139
The Contractile Process in Smooth Muscle 140
Membrane Potentials and Action Potentials in Smooth Muscle 141
Excitation-Contraction Coupling—Role of Calcium Ions 143
Neuromuscular Junctions of Smooth Muscle 144
Smooth Muscle Contraction Wit[illegible] Action Potentials—Effect of [illegible] Tissue Factors and Hormone
Mechanical Characteristics [illegible]
Muscle Contraction .....

IV THE HEART 149

13 HEART MUSCLE; THE HEART AS A PUMP 150
Physiology of Cardiac Muscle 150
Physiologic Anatomy of Cardiac Muscle 150
Action Potentials in Cardiac Muscle .. 151
Contraction of Cardiac Muscle 152
The Cardiac Cycle 153
Systole and Diastole 153
Relationship of the Electrocardiogram to the Cardiac Cycle 154
Function of the Ventricles as Pumps 155
Function of the Valves 155
The Aortic Pressure Curve 156
Relationship of the Heart Sounds to Heart Pumping 156
Work Output of the Heart 157
Energy for Cardiac Contraction 157
Regulation of Cardiac Function 157
Intrinsic Regulation of Cardiac Pumping—The Frank-Starling Law of the Heart 157
Control of the Heart by Parasympathetic and Sympathetic Nerves 160
Effect of Heart Debility on Cardiac Function—The Hypoeffective Heart . 161
Effect of Chronic Increase in Heart Work Load—The Hypereffective Heart 161
Effect of Various Ions on Heart Function 161
Effect of Temperature on the Heart .. 162
The Heart-Lung Preparation 162
Assessment of Contractility 163

14 RHYTHMIC EXCITATION OF THE HEART 165
The Special Excitatory and Conductive System of the Heart 165
The Sinoatrial Node 165
Internodal Pathways and Transmission of the Cardiac Impulse Through the Atria 167
The Atrioventricular (A-V) Node and the Purkinje System 167
Transmission in the Purkinje System 168
Transmission of the Cardiac Impulse in the Ventricular Muscle 168
Summary of the Spread of the Cardiac Impulse Through the Heart .. 169
Control of Excitation and Conduction in the Heart 169
The S-A Node as the Pacemaker of the Heart 169
Role of the Purkinje System in Causing Synchronous Contraction of the Ventricular Muscle 170
Control of Heart Rhythmicity and Conduction by the Autonomic Nerves 170
Abnormal Rhythms of the Heart 171
Heart Block 171
The Phenomenon of "Re-Entry" and Abnormal Rhythms 172
Premature Contraction—Ectopic Foci 175
Cardiac Tachycardias 175
Cardiac Arrest 175

15 THE NORMAL ELECTROCARDIOGRAM 177
Characteristics of the Normal Electrocardiogram 177
Depolarization Waves Versus Repolarization Waves 177
Relationship of Atrial and Ventricular Contraction to the Waves of the Electrocardiogram 178
Voltage and Tima Calibration of the Electrocardiogram 178
Methods for Recording Electrocardiograms 179
The Pen Recorder 179
Recording Electrocardiograms with the Oscilloscope 179
Flow of Current Around the Heart During the Cardiac Cycla 179
Recording Electrical Potentials from a Partially Depolarized Mass of Syncytial Cardiac Muscle 179
Flow of Electrical Currents Around the Heart in the Chest 180
Electrocardiographic Leads 180
The Three Bipolar Limb Leads 180
Chest Leads (Precordial Leads) 181
Augmented Unipolar Limb Leads 182

16 ELECTROCARDIOGRAPHIC INTERPRETATION IN CARDIAC MUSCLE ABNORMALITIES—VECTORIAL ANALYSIS 183
Principles of Vectorial Analysis of Electrocardiograms 183
Use of Vectors to Represent Electrical Potentials 183
Denoting the Direction of a Vector in Terms of Degrees 184
"Axis" of Each of the Bipolar and Unipolar Limb Leads 184
Vectorial Analysis of Potentials Recorded in Different Leads 184
Vectorial Analysis of the Normal Electrocardiogram 185
Vectors Occurring During Depolarization of the Ventricles—the QRS Complex 185

The Electrocardiogram During Repolarization—The T Wave 187
Depolarization of the Atria—The P Wave 187
The Vectorcardiogram 188
The Mean Electrical Axis of the Ventricle 188
Determining the Electrical Axis from Standard Lead Electrocardiograms ... 188
Abnormal Ventricular Conditions That Cause Axis Deviation 189
Conditions that Cause Abnormal Voltages of the QRS Complex 191
Increased Voltage in the Standard Bipolar Limb Leads 191
Decreased Voltage of the Electrocardiogram 191
Prolonged and Bizarre Patterns of the QRS Complex 191
Prolonged QRS Complex as a Result of Cardiac Hypertrophy or Dilatation 191
Prolonged QRS Complex Resulting from Purkinje System Blocks 192
Conditions Causing Bizarre QRS Complexes 192
Current of Injury 192
Effect of Current of Injury on the QRS Complex 192
The J Point—The Zero Reference Potential of the Electrocardiogram 193
Coronary Ischemia as a Cause of Current of Injury 194
Abnormalities in the T Wave 195
Effect of Slow Conduction of the Depolarization Wave on the T Wave 195
Prolonged Depolarization in Portions of the Ventricular Muscle as a Cause of Abnormalities in the T Wave 196

17 ELECTROCARDIOGRAPHIC INTERPRETATION OF CARDIAC ARRHYTHMIAS 197
Abnormal Sinus Rhythms 197
Tachycardia 197
Bradycardia 197
Sinus Arrhythmia 197
Abnormal Rhythms Resulting from Impulse Conduction Block 198
Sinoatrial Block 198
Atrioventricular Block 198
Incomplete Intraventricular Block—Electrical Alternans 199
Premature Contractions 199
Premature Atrial Contractions 200
A-V Nodal or A-V Bundle Premature Contractions 200
Premature Ventricular Contractions (PVCs) 200
Paroxysmal Tachycardia 201
Atrial Paroxysmal Tachycardia 201
Ventricular Paroxysmal Tachycardia 202
Flutter and Fibrillation 202
Atrial Flutter 202
Atrial Fibrillation 202
Ventricular Fibrillation 203

V THE CIRCULATION 205

18 PHYSICS OF BLOOD, BLOOD FLOW, AND PRESSURE: HEMODYNAMICS 206
The Circulatory System as a "Circuit" 206
The Physical Characteristics of Blood 206
The Hematocrit 206
Plasma 208
Interrelationships Among Pressure, Flow, and Resistance 208
Blood Flow 208
Blood Pressure 210
Resistance of Blood Flow 211
Vascular Distensibility 214
Vascular Compliance (Or Capacitance) 214
Volume-Pressure Curves of the Arterial and Venous Circulations 214
"Mean Circulatory Filling Pressure" and Volume-Pressure Curves of the Entire Circulatory System 215
The Mean Circulatory Filling Pressure 215
Volume-Pressure Curves of the Entire Circulation 215
Delayed Compliance (Stress-Relaxation) of Vessels 216

19 THE SYSTEMIC CIRCULATION 218
Physical Characteristics of the Systemic Circulation 218
Function of the Large Arteries 220
The Small Arteries, Arterioles, and Capillaries 220
Exchange of Fluid and Nutrients Through the Capillary Membrane 220
The Veins and Their Functions 221
Venous Pressures—Right Atrial Pressure (Central Venous Pressure) and Peripheral Pressures 221
Blood Reservoir Function of the Veins 224
Pressure Pulses in the Arteries 225
Factors That Affect the Pulse Pressure 225

Abnormal Pressure Pulse Contours... 226
Transmission of the Pressure Pulse to the Periphery 227
The Radial Pulse 228
Pressure Pulses in the Veins 228

20 LOCAL CONTROL OF BLOOD FLOW BY THE TISSUES, AND NERVOUS AND HUMORAL REGULATION..... 230
Local Control of Blood Flow in Response to Tissue Need..... 230
Mechanisms of Blood Flow Control ... 231
Long-Term Blood Flow Regulation..... 235
Nervous Regulation of the Circulation . 237
The Autonomic Nervous System 237
"Patterns" of Circulatory Responses Elicited by Different Central Nervous System Centers..... 240
Reflex Regulation of the Circulation.. 241
Humoral Regulation of the Circulation.... 241

21 ARTERIAL PRESSURE REGULATION: I. RAPID PRESSURE CONTROL BY NERVOUS REFLEXES AND OTHER MECHANISMS 244
Normal Arterial Pressures 244
The Mean Arterial Pressure 244
Relationship of Arterial Pressure to Cardiac Output and Total Peripheral Resistance 246
The Overall System for Arterial Pressure Regulation 246
Rapidly Acting Nervous Mechanisms for Arterial Pressure Control 247
The Arterial Baroreceptor Control System—Baroreceptor Reflexes... .. . 247
Control of Arterial Pressure by the Carotid and Aortic Chemoreceptors—Effect of Oxygen Lack on Arterial Pressure 250
Atrial and Pulmonary Artery Reflexes That Help Regulate Arterial Pressure and Other Circulatory Factors. 250
The CNS Ischemic Response—Control of Arterial Pressure by the Vasomotor Center in Response to Diminished Brain Blood Flow. 250
Special Features of Nervous Control of Arterial Pressure..... 251
Participation of the Veins in Nervous Regulation of Cardiac Output and Arterial Pressure..... 251
Role of the Skeletal Nerves and Skeletal Muscles in Increasing Cardiac Output and Arterial Pressure. 252
Respiratory Waves in the Arterial Pressure..... 252
Arterial Pressure Vasomotor Waves—Oscillation of the Pressure Reflex Control Systems..... 252
Hormonal Mechanisms for Rapid Control of Arterial Pressure..... 25
The Norephinephrine-Epinephrine Vasoconstrictor Mechanism..... 25
Role of Vasopressin in Rapid Control of Arterial Pressure..... 25
The Renin-Angiotensin Vasoconstriction Mechanism for Control of Arterial Pressure 25
Two Intrinsic Circulatory Mechanisms for Rapid Arterial Pressure Regulation 25

22 ARTERIAL PRESSURE REGULATION: II. LONG-TERM CONTROL BY THE RENAL–BODY FLUID MECHANISM AND THE RENIN–ANGIOTENSIN SYSTEM: MECHANISMS OF HYPERTENSION..... 25
The Renal–Body Fluid System for Arterial Pressure Control 25
Role of the Renin-Angiotensin System and of Aldosterone in Long-Term Arterial Pressura Regulation 26
Hypertension (High Blood Pressure)..... 26
Volume-loading Hypertension 26
Vasoconstrictor Hypertension..... 26
Types of Hypertension with Both Volume-Loading and Vasoconstrictor Components. 26
Essential Hypertension... 26
Effects of Hypertension on tha Body.. 27

23 CARDIAC OUTPUT, VENOUS RETURN, AND THEIR REGULATION 27
Normal Values for Cardiac Output at Rest and During Activity 27
Regulation of Cardiac Output 27
Role of the Heart in Cardiac Output Regulation—Its "Permissive" Role.... 27
Role of Total Peripheral Resistance in Determining Normal Venous Return and Cardiac Output. 27
Effect of Arteriovenous Fistulae on Cardiac Output..... 27
Some Special Problems in Cardiac Output Regulation 27
Role of Blood Volume and the "Mean Systemic Filling Pressure" in Cardiac Output Regulation 27
Regulation of Cardiac Output in Heavy Exercise, Which Requires Simultaneous Peripheral and Cardiac Adjustments 27
Abnormally Low and Abnormally High Cardiac Outputs 27
Graphical Analysis of Cardiac Output Regulation 27
Cardiac Output Curves 27
Venous Return Curves 28
Analysis of Cardiac Output and Right Atrial Pressure, Using Cardiac Output and Venous Return Curves ... 28

Methods of Measuring Cardiac Output 284
The Indicator Dilution Method 284

24 THE PULMONARY CIRCULATION..... 287
Physiologic Anatomy of the Pulmonary Circulatory System..... 287
Pressures in the Pulmonary System 288
The Blood Volume of the Lungs..... 288
Blood Flow Through the Lungs and Its Distribution 289
Effect of Hydrostatic Pressure Gradients in the Lungs on Regional Pulmonary Blood Flow 289
Effect of Increased Cardiac Output on the Pulmonary Circulation During Heavy Exercise..... 290
Function of the Pulmonary Circulation When the Left Atrial Pressure Rises as a Result of Left Heart Failure 291
Pulmonary Capillary Dynamics..... 291
Capillary Exchange of Fluid in the Lungs..... 292
Pathological Conditions That Obstruct Blood Flow Through the Lungs..... 293

25 THE CORONARY CIRCULATION AND ISCHEMIC HEART DISEASE 295
Normal Coronary Blood Flow and Its Variations..... 295
Physiologic Anatomy of the Coronary Blood Supply..... 295
Normal Coronary Blood Flow..... 295
Control of Coronary Blood Flow..... 296
Local Metabolism as the Primary Controller of Coronary Flow..... 296
Nervous Control of Coronary Blood Flow 298
The Substrates of Cardiac Metabolism..... 299
Ischemic Heart Disease..... 299
Myocardial Infarction..... 300
Causes of Death Following Acute Coronary Occlusion 300
The Stages of Recovery from Acute Myocardial Infarction..... 301
Function of the Heart Following Recovery from Myocardial Infarction..... 302
Pain in Coronary Disease 302
Angina Pectoris 302
Surgical Treatment of Coronary Disease 303
Measurement of Human Coronary Blood Flow 303

26 CARDIAC FAILURE 305
Dynamics of the Circulation in Cardiac Failure..... 305
Acute Effects of Moderate Cardiac Failure..... 305
The Chronic Stage of Failure..... 306
Summary of the Changes That Occur Following Acute Cardiac Failure—"Compensated Heart Failure" ... 307
Dynamics of Severe Cardiac Failure—"Decompensated Heart Failure"..... 308
Unilateral Cardiac Failure..... 309
Unilateral Left Heart Failure 309
Unilateral Right Heart Failure 309
"High Cardiac Output Failure"—Overloading of the Heart..... 310
Low Cardiac Output in Heart Failure—"Cardiogenic Shock" 310
Edema in Patients with Cardiac Failure .. 311
Physiological Classification of Cardiac Failure..... 312
Cardiac Reserve 312
Appendix 313

27 HEART SOUNDS; DYNAMICS OF VALVULAR AND CONGENITAL HEART DEFECTS..... 316
The Heart Sounds.. 316
Normal Heart Sounds 316
Areas for Auscultation of Normal Heart Sounds 317
The Phonocardiogram 318
Valvular Lesions 318
Abnormal Heart Sounds Caused by Valvular Lesions 318
Abnormal Circulatory Dynamics in Valvular Heart Disease 319
Dynamics of the Circulation in Aortic Stenosis and Aortic Regurgitation 319
Dynamics of Mitral Stenosis and Mitral Regurgitation..... 320
Circulatory Dynamics During Exercise in Patients with Valvular Lesions 321
Abnormal Circulatory Dynamics in Congenital Heart Defects 321
Patent Ductus Arteriosus—A Left-to-Right Shunt 322
Interventricular Septal Defect—A Left-to-Right Shunt 323
Tetralogy of Fallot—A Right-to-Left Shunt 323
Pulmonary Stenosis..... 324
Use of Extracorporeal Circulation During Cardiac Surgery 324
Hypertrophy of the Heart in Valvular and Congenital Heart Disease 324

28 CIRCULATORY SHOCK AND PHYSIOLOGY OF ITS TREATMENT 326
The Physiological Causes of Shock..... 326
Shock Caused by Hypovolemia—Hemorrhagic Shock 327
Relationship of Bleeding Volume to Cardiac Output and Arterial Pressure 327
Nonprogressive and Progressive Hemorrhagic Shock 327

Irreversible Shock 331
Hypovolemic Shock Caused by Plasma Loss 331
Hypovolemic Shock Caused by Trauma 332
Neurogenic Shock—Increased Vascular Capacity 332
Anaphylactic Shock 332
Septic Shock 333
Effects of Shock on the Body 333
Physiology of Treatment in Shock 334
Replacement Therapy 334
Treatment of Shock with Sympathomimetic and Sympatholytic Drugs 334
Other Therapy 334
Circulatory Arrest 335
Effect of Circulatory Arrest on the *Brain* 335

29 MUSCLE BLOOD FLOW DURING EXERCISE; CEREBRAL, SPLANCHNIC, AND SKIN BLOOD FLOWS 336
Blood Flow Through Skeletal Muscles and Its Regulation in Exercise 336
Rate of Blood Flow Through the Muscles 336
Control of Blood Flow Through the Skeletal Muscles 33
Circulatory Readjustments During Exercise 33
The Cerebral Circulation 33
Normal Rate of Cerebral Blood Flow 33
Regulation of Cerebral Blood Flow 33
The Splanchnic Circulation 34
Blood Flow Through the Liver 34
Blood Flow Through the Intestinal Vessels 34
Portal Venous Pressure 34
The Splenic Circulation 34
Circulation in the Skin 34
Physiologic Anatomy of the Cutaneous Circulation 34
Regulation of Blood Flow in the *Skin* 34
Color of the Skin in Relation to Skin Temperature 34
Physiology of Vascular Diseases of the Limbs 34
Raynaud's Disease 34
Peripheral Arteriosclerosis 34

IV THE BODY FLUIDS AND KIDNEYS 34'

30 CAPILLARY DYNAMICS, AND EXCHANGE OF FLUID BETWEEN THE BLOOD AND INTERSTITIAL FLUID 348
Structure of the Capillary System 348
Flow of Blood in the Capillaries—Vasomotion 349
Average Function of the Capillary System 349
Exchange of Nutrients and Other Substances Between the Blood and Interstitial Fluid 350
Diffusion Through the Capillary Membrane 350
The Interstitium and the Interstitial Fluid 351
Distribution of Fluid Volume Between the Plasma and Interstitial Fluid 352
Capillary Pressure 353
Interstitial Fluid Pressure—Intragel Pressure and Free Fluid Pressure 354
Plasma Colloid Osmotic Pressure 356
Interstitial Fluid Colloid Osmotic Pressure 357
Exchange of Fluid Volume Through the Capillary Membrane 357
The Starling Equilibrium for Capillary Exchange 358

31 THE LYMPHATIC SYSTEM, INTERSTITIAL FLUID DYNAMICS, EDEMA, AND PULMONARY FLUID 361
The Lymphatic System 36
The Lymph Channels of the Body 36
Formation of Lymph 36
Total Rate of Lymph Flow 36
Control of Interstitial Fluid Protein Concentration 36
Regulation of Interstitial Fluid Protein by Lymphatic Pumping 36
Control of Interstitial Fluid Pressure 36
Edema 36
Pressure-Volume Curve of the Interstitial Fluid Spaces 36
Positive Interstitial Fluid Pressure as the Physical Basis for Edema 36
The Concept of an "Edema Safety Factor" 36
Edema Resulting from Abnormal Capillary Dynamics 36
Edema Caused by Kidney Retention of Fluid 37
The Presence and Importance of Gel in the Interstitial Spaces 37
Pulmonary Interstitial Fluid Dynamics 37
Pulmonary Edema 37

32 THE SPECIAL FLUID SYSTEMS OF THE BODY—CEREBROSPINAL, OCULAR, PLEURAL, PERICARDIAL, PERITONEAL, AND SYNOVIAL 37
The Cerebrospinal Fluid System 37
Cushioning Function of the Cerebrospinal Fluid 37

Formation, Flow, and Absorption of Cerebrospinal Fluid 375
Cerebrospinal Fluid Pressure 376
Obstruction to the Flow of Cerebrospinal Fluid 376
The Blood-Cerebrospinal Fluid and Blood-Brain Barriers 377
The Intraocular Fluid 377
Formation of Aqueous Humor by the Ciliary Body 378
Outflow of Aqueous Humor from the Eye 378
Intraocular Pressure 378
Fluid Circulation in the Potential Spaces of the Body 379
Fluid Exchange Between the Capillaries and the Potential Spaces 379
The Pleural Cavity 380
The Pericardial Cavity 380
The Peritoneal Cavity 380
The Synovial Cavities 381

33 PARTITION OF THE BODY FLUIDS: OSMOTIC EQUILIBRIA BETWEEN EXTRACELLULAR AND INTRACELLULAR FLUIDS 382
Total Body Water 382
Intake Versus Output of Water 382
Body Fluid Compartments 383
The Intracellular Fluid Compartment 383
The Extracellular Fluid Compartment 383
Blood Volume 383
Measurement of Body Fluid Volumes 384
The Dilution Principle for Measuring Fluid Volumes 384
Determination of Blood Volume 385
Measurement of the Extracellular Fluid Volume 386
Measurement of Total Body Water 386
Calculation of Interstitial Fluid Volume 386
Constituents of Extracellular and Intracellular Fluids 386
Osmotic Equilibria and Fluid Shifts Between the Extracellular and Intracellular Fluids 387
Basic Principles of Osmosis and Osmotic Pressure 387
Osmolality of the Body Fluids 388
Maintenance of Osmotic Equilibrium Between Extracellular and Intracellular Fluids 389
Changes in the Volumes and Osmolalities of the Extracellular and Intracellular Fluid Compartments in Abnormal States 390
Calculation of Fluid Shifts Between the Extracellular and Intracellular Fluid Compartments 390
Effect of Adding Water to the Extracellular Fluid 390
Effect of Dehydration 391
Effect of Adding Saline Solution to the Extracellular Fluid 391
Glucose and Other Solutions Administered for Nutritive Purposes 391

34 FORMATION OF URINE BY THE KIDNEY: GLOMERULAR FILTRATION, TUBULAR FUNCTION, AND PLASMA CLEARANCE 393
Physiologic Anatomy of the Kidney 393
Basic Theory of Nephron Function 395
Renal Blood Flow Through the Kidneys 395
Blood Flow Through the Kidneys 395
Pressures in the Renal Circulation 396
"Intrarenal Pressure" and Renal Interstitial Fluid Pressure 396
Function of the Peritubular Capillaries 396
Glomerular Filtration and the Glomerular Filtrate 397
The Glomerular Filtration Rate 398
Dynamics of Glomerular Filtration 398
Factors That Affect the Glomerular Filtration Rate 399
Reabsorption and Secretion in the Tubules 400
Basic Mechanisms of Absorption and Secretion in the Tubules 400
Absorptive Capabilities of Different Tubule Segments 402
Reabsorption and Secretion of Individual Substances in Different Segments of the Tubules 404
Concentrations of Different Substances at Different Points in the Tubules 406
The Concept of "Plasma Clearance" 407
Inulin Clearance as a Measure of Glomerular Filtration Rate 407
Para-Aminohippuric Acid (PAH) Clearance as a Measure of Plasma Flow Through the Kidneys 408
Calculating the Filtration Fraction from Plasma Clearances 408
Effect of "Tubular Load" and "Tubular Transport Maximum" on Urine Constituents 408

35 RENAL MECHANISMS FOR (1) CONTROLLING GLOMERULAR FILTRATION, (2) EXCRETING DILUTE OR CONCENTRATED URINE, AND (3) EXCRETING UREA, SODIUM, POTASSIUM, AND FLUID VOLUME 410
Autoregulation of Glomerular Filtration Rate 410
Mechanism of Autoregulation of Glomerular Filtration Rate—Tubuloglomerular Feedback 411
Autoregulation of Renal Blood Flow 413
The Mechanism for Excreting Excess Water: Excretion of a Dilute Urine 413

The Mechanism for Excreting Excess Solutes: The Countercurrent Mechanism for Excreting a Concentrated Urine 414
Osmolar Clearance: Free Water Clearance 417
Urea Excretion 417
Sodium Excretion 419
Potassium Excretion 420
Fluid Volume Excretion 421
Summary of the Control of Fluid Volume Excretion 424

36 REGULATION OF BLOOD VOLUME, EXTRACELLULAR FLUID VOLUME, AND EXTRACELLULAR FLUID COMPOSITION BY THE KIDNEYS AND BY THE THIRST MECHANISM 425
Control of Blood Volume 425
Basic Mechanism for Blood Volume Control 426
Role of the Volume Receptors in Blood Volume Control 426
Other Factors That Help Control Blood Volume 427
Control of Extracellular Fluid Volume 428
Control of Extracellular Fluid Sodium Concentration and Extracellular Fluid Osmolality 428
The Osmosodium Receptor–Antidiuretic Hormone Feedback Control System 429
Thirst, and Its Role in Controlling Sodium Concentration and Extracellular Fluid Osmolality 431
Combined Roles of the Antidiuretic and Thirst Mechanisms for Control of Extracellular Fluid Sodium Concentration and Osmolality 432
Effect of Aldosterone on Sodium Concentration 433
Control of Sodium Intake—Appetite and Craving for Salt 433
Regulation of Potassium Ion Concentration 433
Other Factors That Affect Potassium Ion Concentration 435
Control of the Extracellular Concentration of Other Ions 435

37 REGULATION OF ACID-BASE BALANCE 438
Defense Against Changes in Hydrogen Ion Concentration 439
Function of Acid-Base Buffers 439
The Bicarbonate Buffer System 439
The Buffer Systems of the Body Fluids 441
The Isohydric Principle 442
Respiratory Regulation of Acid-Base Balance 442
Renal Regulation of Hydrogen Ion Concentration 444
Tubular Secretion of Hydrogen Ions 444
Renal Correction of Acidosis—Increase of Bicarbonate Ions in the Extracellular Fluid 445
Renal Correction of Alkalosis—Decrease of Bicarbonate Ions in the Extracellular Fluid 446
Combination of Excess Hydrogen Ions with Tubular Buffers and Transport into the Urine 446
Rapidity of Acid-Base Regulation by the Kidneys 447
Renal Regulation of Plasma Chloride Concentration—The Chloride to Bicarbonate Ratio 448
Clinical Abnormalities of Acid-Base Balance 448
Respiratory Acidosis and Alkalosis 448
Metabolic Acidosis and Alkalosis 448
Effects of Acidosis and Alkalosis on the Body 449
Respiratory Compensation of Metabolic Acidosis or Alkalosis 449
Renal Compensation of Respiratory Acidosis or Alkalosis 449
Physiology of Treatment in Acidosis or Alkalosis 450
Clinical Measurements and Analysis of Acid-Base Abnormalities 450

38 RENAL DISEASE, DIURESIS, AND MICTURITION 452
Renal Disease 452
Acute Renal Failure 452
Chronic Renal Failure—Decrease in Number of Functional Nephrons 453
Hypertensive Kidney Disease 457
The Nephrotic Syndrome—Increased Glomerular Permeability 457
Specific Tubular Disorders 458
Renal Function Tests 458
Diuretics and Mechanisms of Their Actions 459
Micturition 460
Physiologic Anatomy and Nervous Connections of the Bladder 460
Transport of Urine Through the Ureters 461
Tone of the Bladder Wall, and the Cystometrogram During Bladder Filling 461
The Micturition Reflex 461
Abnormalities of Micturition 462

VII RESPIRATION 465

39 PULMONARY VENTILATION 466
Mechanics of Pulmonary Ventilation 466
Basic Mechanisms of Lung Expansion and Contraction 466
Respiratory Pressures 467
Expansibility of the Lungs and Thorax: "Compliance" 468
The "Work" of Breathing..... 469
The Pulmonary Volumes and Capacities.. 470
Recording Changes in Pulmonary Volume—Spirometry..... 470
The Pulmonary "Volumes"..... 470
The Pulmonary "Capacities"..... 470
Significance of the Pulmonary Volumes and Capacities..... 471
Abbreviations and Symbols Used in Pulmonary Function Studies..... 471
Determination of Functional Residual Capacity—The Helium Dilution Method 472
The Minute Respiratory Volume—Respiratory Rate Times Tidal Volume..... 472
Alveolar Ventilation..... 472
Rate of Alveolar Ventilation..... 473
Functions of the Respiratory Passageways..... 474
The Trachea, Bronchi, and Bronchioles..... 474
Maximum Expiratory Flow 476
Forced Expiratory Vital Capacity and Forced Expiratory Volume..... 477
Respiratory Functions of the Nose 477
Vocalization..... 477
Artificial Respiration 478
Mouth-to-Mouth Breathing 478
Mechanical Methods of Artificial Respiration..... 479

40 PHYSICAL PRINCIPLES OF GASEOUS EXCHANGE; DIFFUSION OF OXYGEN AND CARBON DIOXIDE THROUGH THE RESPIRATORY MEMBRANE 481
Physics of Diffusion and Gas Pressures ... 481
The Molecular Basis of Gaseous Diffusion 481
Gas Pressures in a Mixture of Gases—Partial Pressures of Individual Gases 482
Pressures of Gases in Water and Tissues..... 482
The Vapor Pressure of Water..... 482
Diffusion of Gases Through Fluids—The Pressure Gradient for Diffusion.. 483
Diffusion of Gases Through Tissues... 483
Composition of Alveolar Air—Its Relation to Atmospheric Air..... 483
Rate at Which Alveolar Air Is Renewed by Atmospheric Air..... 484
Oxygen Concentration and Partial Pressure in the Alveoli 485
CO_2 Concentration and Partial Pressure in the Alveoli 485
Expired Air 485
Diffusion of Gases Through the Respiratory Membrane 486
Factors That Affect Rate of Gas Diffusion Through the Respiratory Membrane..... 488
Diffusing Capacity of the Respiratory Membrane..... 489
Effect of the Ventilation-Perfusion Ratio on Alveolar Gas Concentration... 490

41 TRANSPORT OF OXYGEN AND CARBON DIOXIDE IN THE BLOOD AND BODY FLUIDS..... 493
Pressures of Oxygen and Carbon Dioxide in the Lungs, Blood, and Tissues..... 493
Uptake of Oxygen by the Pulmonary Blood 493
Transport of Oxygen in the Arterial Blood..... 494
Diffusion of Oxygen from the Capillaries to the Interstitial Fluid ... 494
Diffusion of Oxygen from the Capillaries to the Cells 495
Diffusion of Carbon Dioxide from the Cells to the Tissue Capillaries, and from the Pulmonary Capillaries to the Alveoli 495
Transport of Oxygen in the Blood..... 496
The Reversible Combination of Oxygen with Hemoglobin 496
The Tissue Oxygen Buffer Function of Hemoglobin..... 497
Shift of the Oxygen-Hemoglobin Dissociation Curve and Its Significance 498
Metabolic Use of Oxygen by the Cells 499
Transport of Oxygen in the Dissolved State..... 499
Combination of Hemoglobin with Carbon Monoxide..... 500
Transport of Carbon Dioxide in the Blood..... 500
Chemical Forms in Which Carbon Dioxide Is Transported 500
The Carbon Dioxide Dissociation Curve..... 501
Effect of the Oxygen-Hemoglobin Reaction on Carbon Dioxide Transport—The Haldane Effect 501
Change in Blood Acidity During Carbon Dioxide Transport..... 502
The Respiratory Exchange Ratio..... 502

42 REGULATION OF RESPIRATION 504
The Respiratory Center 504
Control of Overall Respiratory Center Activity 506
Chemical Control of Respiration 506
Direct Chemical Control of Respiratory Center Activity by Carbon Dioxide and Hydrogen Ions ... 506
The Peripheral Chemoreceptor System for Control of Respiratory Activity—Role of Oxygen in Respiratory Control 508
Composite Effects of P_{CO_2}, pH, and P_{O_2} on Respiratory Activity 511
Regulation of Respiration During Exercise 511
Other Factors That Affect Respiration 513
Abnormalities of Respiratory Control 514
Respiratory Center Depression 514
Periodic Breathing 514

43 RESPIRATORY INSUFFICIENCY—PATHOPHYSIOLOGY, DIAGNOSIS, OXYGEN THERAPY 516
Additional Methods for Studying Respiratory Abnormalities 516
Study of Blood Gases and pH 517
Physiologic Types of Respiratory Insufficiency 517
Abnormalities That Cause Alveolar Hypoventilation 518
Diseases That Decrease Lung Diffusing Capacity 518
Abnormalities of Oxygen Transport from the Lungs to the Tissues 518
Physiologic Peculiarities of Specific Pulmonary Abnormalities 518
Chronic Pulmonary Emphysema 518
Pneumonia 520
Atelectasis 520
Asthma 521
Tuberculosis 521
Hypoxia 522
Cyanosis 522
Dyspnea 522
Hypercapnia 523
Effects of Hypercapnia on the Body ... 523
Oxygen Therapy in the Different Types of Hypoxia 523
Danger of Hypercapnia During Oxygen Therapy 524
Absorption of Entrapped Air 525

VIII AVIATION, SPACE, AND DEEP SEA DIVING PHYSIOLOGY 527

44 AVIATION, HIGH ALTITUDE, AND SPACE PHYSIOLOGY 528
Effects of Low Oxygen Pressure on the Body 528
Alveolar P_{O_2} at Different Elevations 528
Effect of Breathing Pure Oxygen on the Alveolar P_{O_2} at Different Altitudes 529
The "Ceiling" When Breathing Air and When Breathing Oxygen in an Unpressurized Airplane 529
Effects of Hypoxia 529
Acclimatization of Low P_{O_2} 530
Natural Acclimatization of Natives Living at High Altitudes 531
Work Capacity at High Altitudes: The Effect of Acclimatization 532
Chronic Mountain Sickness 532
Acute Mountain Sickness and High Altitude Pulmonary Edema 532
Effects of Acceleratary Forces on the Body in Aviation and Space Physiology 532
Centrifugal Acceleratory Forces 532
Effects of Linear Acceleratory Forces on the Body 534
Perception of Equilibrium and Turning in Blind Flying 534
Problems of Temperature in Aviation and Space Physiology 534
Radiation at High Altitudes and in Space 535
"Artificial Climate" in the Sealed Spacecraft 535
Weightlessness in Space 536

45 PHYSIOLOGY OF DEEP SEA DIVING AND OTHER HYPERBARIC CONDITIONS 537
Effect of High Partial Pressures of Gases on the Body 537
Decompression of the Diver After Exposure to High Pressures 539
Some Physical Problems of Diving 541
Scuba Diving (Self-Contained Underwater Breathing Apparatus) 542
Special Physiological Problems in Submarines 543
Hyperbaric Oxygen Therapy ..... 54[illegible]

IX THE NERVOUS SYSTEM 545

46 ORGANIZATION OF THE NERVOUS SYSTEM; BASIC FUNCTIONS OF SYNAPSES 546
General Design of the Nervous System 546
The Sensory Division—Sensory Receptors 546

The Motor Division—The Effectors ... 547
Processing of Information—"Integrative" Function of the Nervous System ... 547
Storage of Information—Memory ... 548
The Three Major Levels of Central Nervous System Function ... 548
Comparison of the Nervous System with an Electronic Computer ... 549
The Central Nervous System Synapses ... 549
Physiologic Anatomy of the Synapse ... 550
Chemical and Physiological Natures of the Transmitter Substances ... 552
Electrical Events During Neuronal Excitation ... 553
Electrical Events in Neuronal Inhibition ... 556
Summation of Postsynaptic Potentials ... 557
Special Functions of Dendrites in Exciting Neurons ... 558
Relation of State of Excitation of the Neuron to the Rate of Firing ... 559
Some Special Characteristics of Synaptic Transmission ... 560

7 NEURONAL MECHANISMS AND CIRCUITS FOR PROCESSING INFORMATION ... 562
Information, Signals, and Impulses ... 562
Transmission of Signals in Nerve Tracts ... 562
Signal Strength ... 562
Transmission and Processing of Signals in Neuronal Pools ... 563
Relaying of Signals Through Neuronal Pools ... 563
Transmission of Spatial Patterns Through Successive Neuronal Pools ... 565
Prolongation of a Signal By a Neuronal Pool—"After-Discharge" ... 567
Rhythmic Signal Output ... 569
Instability and Stability of Neuronal Circuits ... 569
Inhibitory Circuits as a Mechanism for Stabilizing Nervous System Function ... 570
Synaptic Fatigue as a Means for Stabilizing the Nervous System ... 570

8 SENSORY RECEPTORS AND THEIR BASIC MECHANISMS OF ACTION ... 572
Types of Sensory Receptors and the Sensory Stimuli They Detect ... 572
Differential Sensitivity of Receptors ... 572
Transduction of Sensory Stimuli into Nerve Impulses ... 573
Local Currents at Nerve Endings—Receptor Potentials ... 573
Adaptation of Receptors ... 575
Psychic Interpretation of Stimulus Strength ... 576
Judgment of Stimulus Strength ... 577
Physiological Classification of Nerve Fibers ... 578

49 SOMATIC SENSATIONS: I. THE MECHANORECEPTIVE SENSATIONS ... 580
Classification of Somatic Senses ... 580
Detection and Transmission of Tactile Sensations ... 580
Detection of Vibration ... 582
Tickling and Itch ... 582
The Dual System for Transmission of Mechanoreceptive Somatic Sensory Signals into the Central Nervous System ... 582
Transmission in the Dorsal Column—Lemniscal System ... 583
Anatomy of the Dorsal Column—Lemniscal System ... 583
The Somatic Sensory Cortex ... 584
Somatic Association Areas ... 586
Characteristics of Transmission in the Dorsal Column—Lemniscal System ... 587
The Position Sense ... 588
Transmission in the Anterolateral System ... 589
Anatomy of the Anterolateral Pathway ... 589
Some Special Aspects of Sensory Function ... 590

50 SOMATIC SENSATIONS: II. PAIN, VISCERAL PAIN, HEADACHE, AND THERMAL SENSATIONS ... 592
The Two Types of Pain and Their Qualities—Acute Pain and Slow Pain ... 592
Methods for Measuring the Perception of Pain ... 592
The Pain Receptors and Their Stimulation ... 593
Rate of Tissue Damage as a Cause of Pain ... 593
The Dual Pathways for Transmission of Pain Signals into the Central Nervous System ... 594
A Pain Control ("Analgesia") System in the Brain and Spinal Cord ... 596
Inhibition of Pain Transmission at the Cord Level by Tactile Signals ... 597
Referred Pain ... 598
Referred Pain Caused by Reflex Muscular Spasm ... 598
Visceral Pain ... 598
Causes of True Visceral Pain ... 598
"Parietal" Pain Caused by Visceral Damage ... 599
Localization of Visceral Pain—The "Visceral" and the "Parietal" Transmission Pathways ... 599
Visceral Pain from Various Organs ... 600
Some Clinical Abnormalities of Pain and Other Sensations ... 601
Hyperalgesia ... 601

The Thalamic Syndrome 601
Herpes Zoster 601
Tic Douloureux 601
The Brown-Séquard Syndrome 601
Headache 602
Headache of Intracranial Origin 602
Extracranial Types of Headache 603
Thermal Sensations 603
Thermal Receptors and Their Excitation 603
Transmission of Thermal Signals in the Nervous System 605

51 MOTOR FUNCTIONS OF THE SPINAL CORD; AND THE CORD REFLEXES 606
Organization of the Spinal Cord for Motor Functions 606
The Muscle Receptors—Muscle Spindles and Golgi Tendon Organs—and their Roles in Muscle Control 607
Receptor Function of the Muscle Spindle 608
The Stretch Reflex (Also Called Myotatic Reflex) 609
Role of the Muscle Spindle in Voluntary Motor Activity 611
Clinical Applications of the Stretch Reflex 612
The Tendon Reflex 612
The Flexor Reflex (The Withdrawal Reflexes) 613
The Crossed Extensor Reflex 614
Reciprocal Inhibition and Reciprocal Innervation 615
Fatigue of Reflexes, Rebound 615
The Reflexes of Posture and Locomotion 615
The Postural and Locomotive Reflexes of the Cord 615
The Scratch Reflex 617
The Spinal Cord Reflexes That Cause Muscle Spasm 617
The Autonomic Reflexes in the Spinal Cord 617
Spinal Cord Transection and Spinal Shock 617

52 MOTOR FUNCTIONS OF THE BRAIN STEM AND BASAL GANGLIA—RETICULAR FORMATION, VESTIBULAR APPARATUS, EQUILIBRIUM, AND BRAIN STEM REFLEXES 619
The Reticular Formation, and Support of the Body Against Gravity 619
Vestibular Sensations and the Maintenance of Equilibrium 621
The Vestibular Apparatus 621
Function of the Utricle (and Probably Saccule) in the Maintenance of Static Equilibrium 623
The Semicircular Canals and Their Detection of Head Rotation 623
Vestibular Postural Reflexes ..... [illegible]
Vestibular Mechanism for Stabilizing the Eyes and for Nystagmus ..... [illegible]
Other Factors Concerned with Equilibrium ..... [illegible]
Functions of the Reticular Formation and Specific Brain Stem Nuclei in Controlling Subconscious, Stereotyped Movements ..... [illegible]
The Basal Ganglia—Their Motor Functions ..... [illegible]
Functions of the Basal Ganglia ..... [illegible]
Clinical Syndromes Resulting from Damage to the Basal Ganglia ..... [illegible]

53 CORTICAL AND CEREBELLAR CONTROL OF MOTOR FUNCTIONS ..... [illegible]
The Motor Cortex—The Primary and Premotor Areas ..... [illegible]
The Topographical Map of the Primary Motor Area ..... [illegible]
Complex Movements Elicited by Stimulating the Premotor Area ..... [illegible]
Fiber Pathways to and from the Motor Cortex ..... [illegible]
The Red Nucleus and the Rubrospinal Tract—Their Relationship to the Pyramidal System ..... [illegible]
The Extrapyramidal System ..... [illegible]
Excitation of the Spinal Cord by the Primary Motor Cortex and the Red Nucleus ..... [illegible]
The Cerebellum and Its Motor Functions ..... [illegible]
The Anatomical Functional Areas of the Cerebellum ..... [illegible]
The Neuronal Circuit of the Cerebellum ..... [illegible]
Function of the Cerebellum in Controlling Movements ..... [illegible]
Function of the Cerebellum with the Spinal Cord and Brain Stem to Control Postural and Equilibrium Movements ..... [illegible]
Function of the Cerebellum in Voluntary Muscle Control ..... [illegible]
Cerebellar Feedback Control of the Distal Limbs Through the Intermediate Cerebellar Zone and Interpositus Nucleus ..... [illegible]
Function of the Large Lateral Zone of the Cerebellar Hemisphere—The "Planning" and "Timing" Functions ..... [illegible]
Clinical Abnormalities of the Cerebellum ..... [illegible]
Sensory Feedback Control of Motor Functions ..... [illegible]
The "Sensory Engram" for Motor Control ..... [illegible]
Establishment of Rapid Motor Patterns ..... [illegible]
Initiation of Voluntary Motor Activity ..... [illegible]

54 THE CEREBRAL CORTEX AND INTELLECTUAL FUNCTIONS OF THE BRAIN 651
Physiologic Anatomy of the Cerebral Cortex 651
Functions of Certain Specific Cortical Areas 652
Specific Functions of the Primary Sensory Areas 652
The Sensory Association Areas 653
Interpretative Function of the Posterior Superior Temporal Lobe—"Wernicke's Area" (a General Interpretative Area) 654
The Prefrontal Areas 656
Thoughts, Consciousness, and Memory 656
Memory and Types of Memory 657
Physiological Basis of Memory 658
Analytical Operations of the Brain 661
Function of the Brain in Communication 662
Function of the Corpus Callosum and Anterior Commissure to Transfer Thoughts, Memories, and Other Information to the Opposite Hemisphere . 662

55 ACTIVATION OF THE BRAIN—THE RETICULAR ACTIVATING SYSTEM AND WAKEFULNESS; ATTENTION AND SEARCHING OF MEMORY; BRAIN WAVES; SLEEP; EPILEPSY 665
The Reticular Activating System and Its Role in Wakefulness 665
Specific Neuronal Systems Associated with the Reticular Formation 667
The Generalized Thalamocortical System—Its Possible Function in Attention and Searching the Memory Store 668
Brain Waves 669
Origin of the Different Types of Brain Waves 669
Effect of Varying Degrees of Cerebral Activity on the Basic Rhythm of the Electroencephalogram 670
Clinical Use of the Electroencephalogram 670
Sleep 670
Slow Wave Sleep 671
REM Sleep (Paradoxical Sleep, Desynchronized Sleep) 672
Basic Theories of Sleep 672
Physiological Effects of Sleep 673
Epilepsy 674
Grand Mal Epilepsy 674
Petit Mal Epilepsy 675
Focal Epilepsy 675

56 BEHAVIORAL FUNCTIONS OF THE BRAIN: THE LIMBIC SYSTEM, ROLE OF THE HYPOTHALAMUS, AND CONTROL OF VEGETATIVE FUNCTIONS OF THE BODY 676
Functional Anatomy of the Limbic System; Its Relation to the Hypothalamus 676
The Hypothalamus, A Major Output Pathway of the Limbic System 677
Vegetative and Endocrine Control Functions of the Hypothalamus 678
Behavioral Functions of the Hypothalamus and Associated Limbic Structures 679
Specific Functions of Other Parts of the Limbic System 681
Functions of the Amygdala 681
Functions of the Hippocampus 682
Function of the Limbic Cortex 683
Function of Specific Chemical Transmitter Systems for Behavior Control 683
Psychosomatic Effects of the Behavior System 684

57 THE AUTONOMIC NERVOUS SYSTEM; THE ADRENAL MEDULLA 686
General Organization of the Autonomic Nervous System 688
Physiologic Anatomy of the Sympathetic Nervous System 686
Physiologic Anatomy of the Parasympathetic Nervous System 687
Basic Characteristics of Sympathetic and Parasympathetic Function 688
Cholinergic and Adrenergic Fibers—Secretion of Acetylcholine or Norepinephrine by the Postganglionic Neurons 688
Receptors of the Effector Organs 689
Excitatory and Inhibitory Actions of Sympathetic and Parasympathetic Stimulation 690
Effects of Sympathetic and Parasympathetic Stimulation on Specific Organs 690
Function of the Adrenal Medullae 692
Relationship of Stimulus Rate to Degree of Sympathetic and Parasympathetic Effect 693
Sympathetic and Parasympathetic "Tone" 693
Denervation Supersensitivity of Sympathetic and Parasympathetic Organs Following Denervation 693
The Autonomic Reflexes 694
Stimulation of Discrete Organs in Some Instances, Mass Stimulation in Other Instances, by the Sympathetic and Parasympathetic Systems 694
"Alarm" or "Stress" Response of the Sympathetic Nervous System 695
Medullary, Pontine, and Mesencephalic Control of the Autonomic Nervous System 695
Pharmacology of the Autonomic Nervous System 696

Drugs That Act on Adrenergic Effector Organs—The Sympathomimetic Drugs 696
Drugs That Act on Cholinergic Effector Organs..... 696
Drugs That Stimulate the Postganglionic Neurons—"Nicotinic Drugs"..... 69

X THE SPECIAL SENSES 69

58 THE EYE: I OPTICS OF VISION... 700
Physical Principles of Optics.. 700
Refraction of Light... 700
Application of Refractive Principles to Lenses 700
Focal Length of a Lens 702
Formation of an Image by a Convex Lens 702
Measurement of the Refractive Power of a Lens—The Diopter 702
The Optics of the Eye 703
The Eye as a Camera..... 703
The Mechanism of Accommodation ... 704
The Pupillary Aperture... 705
Errors of Refraction 705
Visual Acuity 707
Determination of Distance of an Object from the Eye—Depth Perception 708
Optical Instruments. 708
The Ophthalmoscope.. 708
The Retinoscope 709

59 THE EYE: II RECEPTOR AND NEURAL FUNCTION OF THE RETINA. 711
Anatomy and Function of the Structural Elements of the Retina 711
Photochemistry of Vision 713
The Rhodopsin-Retinal Visual Cycle, and Excitation of the Rods 713
Photochemistry of Color Vision by the Cones 716
Automatic Regulation of Retinal Sensitivity—Dark and Light Adaptation 716
Color Vision 717
The Tri-Color Mechanism of Color Perception 718
Color Blindness 718
Neural Function of the Retina. .. . 720
Neural Organization of the Retina. .. 720
Stimulation of the Rods and Cones ... 720
Stimulation of the Bipolar Cells . .. 721
Stimulation and Function of the Horizontal Cells 721
Stimulation and Function of the Amacrine Cells 721
Excitation of the Ganglion Cells 721
Different Types of Signals Transmitted by the Ganglion Cells Through the Optic Nerve..... 722

60 THE EYE. III. CENTRAL NEUROPHYSIOLOGY OF VISION 72
The Visual Pathway 72
Three Different Types of Retinal Ganglial Cells and Their Projections 72
Function of the Primary Visual Cortex... 72
Detection of Lines and Borders by the Primary Visual Cortex 72
Analysis of Color by the Visual Cortex 72
Perception of Luminosity 72
Effect of Removing the Primary Visual Cortex... 72
Transmission of Visual Information into Other Regions of the Cerebral Cortex 72
The Fields of Vision; Perimetry 72
Eye Movements and Their Control 72
Fixation Movements of the Eyes 72
Fusion of the Visual Images 73
Autonomic Control of Accommodation and Pupillary Aperture. 73
Control of Accommodation (Focusing the Eyes).. 73
Control of Pupillary Diameter..... 73

61 THE SENSE OF HEARING 73
The Tympanic Membrane and the Ossicular System.. 73
Conduction of Sound from the Tympanic Membrane to the Cochlea..... 73
Transmission of Sound through Bone 73
The Cochlea 73
Functional Anatomy of the Cochlea .. 73
Transmission of Sound Waves in the Cochlea—The "Traveling Wave" 73
Function of the Organ of Corti 73
Determination of Sound Frequency—The "Place" Principle..... 73
Determination of Loudness 73
General Auditory Mechanisms 74
The Auditory Pathway 74
Function of the Cerebral Cortex in Hearing..... 74
Discrimination of Direction from Which Sound Emanates 74
Centrifugal Conduction of Impulses from the Central Nervous System... 74
Hearing Abnormalities 74

Types of Deafness 743
62 THE CHEMICAL SENSES—TASTE AND SMELL 745
The Sense of Taste 745
The Primary Sensations of Taste 745
The Taste Bud and Its Function 746
Transmission of Taste Signals into the Central Nervous System 747
Special Attributes of the Taste Sense 748
The Sense of Smell 748
The Olfactory Membrane 749
Stimulation of the Olfactory Cells 749
Transmission of Smell Signals into the Central Nervous System 750

XI THE GASTROINTESTINAL TRACT 753

63 MOVEMENT OF FOOD THROUGH THE ALIMENTARY TRACT 754
General Principles of Gastrointestinal Motility 754
Characteristics of the Gastrointestinal Wall 754
Innervation of the Gut—The Enteric Nervous System 756
Hormonal Control of Gastrointestinal Motility 758
Functional Types of Movement in the Gastrointestinal Tract 758
The Mixing Movements 758
The Propulsive Movements—Peristalsis 758
Ingestion of Food 759
Mastication (Chewing) 759
Swallowing (Deglutition) 760
Function of the Lower Esophageal Sphincter (Gastroesophageal Sphincter) 761
Motor Functions of the Stomach 762
Storage Function of the Stomach 762
Mixing and Propulsion of Food in the Stomach—The Basic Electrical Rhythm of the Stomach 762
Emptying of the Stomach 763
Movements of the Small Intestine 765
Mixing Contractions (Segmentation Contractions) 765
Propulsive Movements 766
Function of the Ileocecal Valve 766
Movements of the Colon 767
Defecation 768
Other Autonomic Reflexes Affecting Bowel Activity 769
64 SECRETORY FUNCTIONS OF THE ALIMENTARY TRACT 770
General Principles of Gastrointestinal Secretion 770
Anatomical Types of Glands 770
Basic Mechanisms of Stimulation of the Gastrointestinal Glands 770
Basic Mechanism of Secretion by Glandular Cells 771
Lubricating and Protective Properties of Mucus and Its Importance in the Gastrointestinal Tract 772
Secretion of Saliva 772
Esophageal Secretion 774
Gastric Secretion 774
Characteristics of the Gastric Secretions 774
Regulation of Gastric Secretion by Nervous and Hormonal Mechanisms 776
Pancreatic Secretion 778
Regulation of Pancreatic Secretion 780
Secretion of Bile and Functions of the Biliary Tree 781
Physiologic Anatomy of Biliary Secretion 781
The Bile Salts and Their Function 782
Secretion of Cholesterol; Gallstone Formation 783
Secretions of the Small Intestine 784
Secretion of Mucus by Brunner's Glands and by Mucous Cells of the Intestinal Surface 784
Secretion of the Intestinal Digestive Juices—The Crypts of Lieberkuhn 784
Regulation of Small Intestinal Secretion 785
Secretions of the Large Intestine 785
65 DIGESTION AND ABSORPTION IN THE GASTROINTESTINAL TRACT 787
Digestion of the Various Foods 787
Digestion of Carbohydrates 787
Digestion of Fats 788
Digestion of Proteins 790
Basic Principles of Gastrointestinal Absorption 790
Anatomical Basis of Absorption 790
Basic Mechanisms of Absorption 792
Absorption in the Small Intestine 792
Absorption of Water 792
Absorption of Ions 792
Absorption of Nutrients 794
Absorption in the Large Intestine: Formation of the Feces 796
66 PHYSIOLOGY OF GASTROINTESTINAL DISORDERS 798
Disorders of Swallowing and of the Esophagus 798
Disorders of the Stomach 798

Peptic Ulcer 799
Disorders of the Small Intestine 801
Disorders of the Large Intestine 802
Constipation 802
Diarrhea 802
Paralysis of Defecation in Spinal Cord Injuries 803
General Disorders of the Gastrointestinal Tract 803
Vomiting 803
Nausea 804
Gastrointestinal Obstruction 804
Gases in the Gastrointestinal Tract and Flatus 805

XII METABOLISM AND TEMPERATURE REGULATION 807

67 METABOLISM OF CARBOHYDRATES AND FORMATION OF ADENOSINE TRIPHOSPHATE 808
Release of Energy from Foods and the Concept of "Free Energy" 808
Role of Adenosine Triphosphate (ATP) in Metabolism 808
Central Role of Glucose in Carbohydrate Metabolism 809
Transport of Glucose Through the Cell Membrane 809
Facilitation of Glucose Transport by Insulin 810
Phosphorylation of Glucose 810
Storage of Glycogen in Liver and Muscle 810
Glycogenesis 810
Glycogenolysis 811
Release of Energy from the Glucose Molecule by the Glycolytic Pathway 811
Glycolysis and the Formation of Pyruvic Acid 811
Conversion of Pyruvic Acid to Acetylcoenzyme A 812
The Citric Acid Cycle 812
Formation of ATP by Oxidative Phosphorylation 813
Summary of ATP Formation During the Breakdown of Glucose 814
Control of Glycolysis and Oxidation by ATP and ADP 815
Anaerobic Release of Energy—"Anaerobic Glycolysis" 815
Release of Energy from Glucose by the Phosphogluconate Pathway 816
Glucose Conversion to Glycogen or Fat 816
ormation of Carbohydrates from Proteins Fats—"Gluconeogenesis" 816
Blood Glucose 817

68 LIPID METABOLISM 818
Transport of Lipids in The Blood 818
Transport from the Gastrointestinal Tract—The Chylomicrons 818
Transport of Fatty Acids in Combination with Albumin—"Free Fatty Acid" 819
The Lipoproteins 819
The Fat Deposits 820
Adipose Tissue 820
The Liver Lipids 820
Use of Triglycerides for Energy, and Formation of Adenosine Triphosphate (ATP) 820
Formation of Acetoacetic Acid in the Liver and Its Transport in the Blood 821
Synthesis of Triglycerides from Carbohydrates 822
Synthesis of Triglycerides from Proteins 823
Regulation of Energy Release from Triglycerides 823
Obesity 824
Phospholipids and Cholesterol 824
Phospholipids 824
Cholesterol 825
Cellular Structural Functions of Phospholipids and Cholesterol—Especially for Membranes 826
Atherosclerosis 826
Experimental Production of Atherosclerosis in Animals 826
Atherosclerosis in the Human Being 827
Summary of Factors Causing Atherosclerosis 827

69 PROTEIN METABOLISM 829
Basic Properties of Proteins 829
The Amino Acids 829
Transport and Storage of Amino Acids 829
The Blood Amino Acids 829
Storage of Amino Acids as Proteins in the Cells 831
Metabolic Roles of the Plasma Proteins 831
Essential and Nonessential Amino Acids 832
Use of Proteins for Energy 832
Obligatory Degradation of Proteins 833
Hormonal Regulation of Protein Metabolism 834

70 THE LIVER AS AN ORGAN 835
Physiologic Anatomy of the Liver 835
Function of the Hepatic Vascular System 836

Metabolic Functions of the Liver 836
Carbohydrate Metabolism 837
Fat Metabolism 837
Protein Metabolism 837
Miscellaneous Metabolic Functions of the Liver 838
Excretion of Bilirubin in the Bile 838

71 ENERGETICS AND METABOLIC RATE 841
Importance of Adenosine Triphosphate (ATP) in Metabolism 841
Phosphocreatine as a Storage Depot for Energy and for Buffering the Concentration of ATP 842
Anaerobic Versus Aerobic Energy 842
Summary of Energy Utilization by the Cells 843
Control of Energy Release in the Cell 843
The Metabolic Rate 844
Measurement of the Metabolic Rate .. 845
Factors That Affect the Metabolic Rate 845
The Basal Metabolic Rate 847

72 BODY TEMPERATURE, TEMPERATURE REGULATION, AND FEVER 849
The Insulator System of the Body 849
Flow of Blood to the Skin and Heat Transfer from the Body Core 850
Balance Between Heat Production and Heat Loss 850
Heat Loss 850
Sweating and Its Regulation by the Autonomic Nervous System 852
Regulation of Body Temperature—The "Hypothalamic Thermostat" 853
Neuronal Effector Mechanisms That Decrease or Increase Body Temperature 854
The Concept of a "Set-Point" for Temperature Control 856
Behavioral Control of Body Temperature 857
Local Skin Temperature Reflexes 857
Abnormalities of Body Temperature Regulation 858
Fever 858
Exposure of the Body to Extreme Cold 859

73 DIETARY BALANCES; REGULATION OF FEEDING; OBESITY AND STARVATION; VITAMINS AND MINERALS 861
Dietary Balances 861
Energy Available in Foods 861
Regulation of Food Intake 863
Neural Centers for Regulation of Food Intake 863
Factors That Regulate Food Intake ... 863
Obesity 865
Treatment of Obesity 866
Inanition 866
Starvation 866
Vitamins 867
Vitamin A 867
Thiamine (Vitamin B_1) 867
Niacin 868
Riboflavin (Vitamin B_2) 868
Vitamin B_{12} 869
Folic Acid (Pteroylglutamic Acid) 869
Pyridoxine (Vitamin B_6) 869
Pantothenic Acid 870
Ascorbic Acid (Vitamin C) 870
Vitamin D 870
Vitamin E 871
Vitamin K 871
Mineral Metabolism 871

XIII ENDOCRINOLOGY AND REPRODUCTION 875

74 INTRODUCTION TO ENDOCRINOLOGY 876
Nature of a Hormone 876
An Overview of the Important Endocrine Glands and Their Hormones 876
Chemistry of the Hormones 878
Hormone Receptors and Their Activation 879
Mechanisms of Hormonal Action 879
The Cyclic AMP Mechanism for Controlling Cell Funtion—A "Second Messenger" for Hormone Mediation .. 880
Action of Steroid Hormones on the Genes to Cause Protein Synthesis 881
Action of the Thyroid Hormones in the Cell Nucleus 881
Measurement of Hormone Concentrations in the Blood 882

75 THE PITUITARY HORMONES AND THEIR CONTROL BY THE HYPOTHALAMUS 884
The Pituitary Gland and Its Relation to the Hypothalamus 884
Cell Types in the Anterior Pituitary Gland 885
Control of Pituitary Secretion by the Hypothalamus 885

The Hypothalamic-Hypophysial Portal System 885
Physiological Functions of Growth Hormone 887
Metabolic Effects of Growth Hormone 887
Stimulation of Cartilage and Bone Growth—Role of the Somatomedins .. 889
Regulation of Growth Hormone Secretion 889
Abnormalities of Growth Hormone Secretion 891
The Posterior Pituitary Gland and Its Relation to the Hypothalamus 893
Chemical Nature of Antidiuretic Hormone (Vasopressin) and Oxytocin 893
Physiological Functions of Antidiuretic Hormone (Vasopressin) .. 893
Oxytocic Hormone 895

76 THE THYROID METABOLIC HORMONES 897
Formation and Secretion of the Thyroid Hormones 897
Iodine Requirements for Formation of Thyroxine 897
The Iodide Pump (Iodide Trapping) 898
Thyroglobulin and Chemistry of Thyroxine and Triiodothyronine Formation 898
Release of Thyroxine and Triiodothyronine from Thyroglobulin 899
Transport of Thyroxine and Triiodothyronine to the Tissues 899
Functions of the Thyroid Hormones in the Tissues 900
General Increase in Metabolic Rate... 900
Effect of Thyroid Hormone on Growth 901
Effects of Thyroid Hormone on Specific Bodily Mechanisms 901
Regulation of Thyroid Hormone Secretion 903
Antithyroid Substances 904
Diseases of the Thyroid 905
Hyperthyroidism 905
Hypothyroidism 906

THE ADRENOCORTICAL HORMONES 909
Chemistry of Adrenocortical Secretion 909
Functions of the Mineralocorticoids—Aldosterone 911
Renal Effects of Aldosterone 911
Effects of Aldosterone on Sweat Glands, Salivary Glands, and Intestinal Absorption 912
Cellular Mechanism of Aldosterone Action 912
Regulation of Aldosterone Secretion .. 913
Functions of the Glucocorticoids 914
Effects of Cortisol on Carbohydrate Metabolism 91
Effects of Cortisol on Protein Metabolism 91
Effects of Cortisol on Fat Metabolism 91
Other Effects of Cortisol 91
Regulation of Cortisol Secretion—Adrenocorticotropic Hormone (ACTH) 91
The Adrenal Androgens 91
Abnormalities of Adrenocortical Secretion 9
Hypoadrenalism—Addison's Disease 9
Hyperodrenalism—Cushing's Syndrome 9
Primary Aldosteronism 9
Adrenogenital Syndrome 9

78 INSULIN, GLUCAGON, AND DIABETES MELLITUS 9
Insulin and Its Metabolic Effects 9
Effect of Insulin on Carbohydrate Metabolism 9
Effect of Insulin on Fat Metabolism .. 9
Effect of Insulin on Protein Metabolism and Growth 9
Control of Insulin Secretion 9
Role of Insulin in "Switching" Between Carbohydrate and Lipid Metabolism 9
Glucagon and Its Functions 9
Effects on Glucose Metabolism 9
Regulation of Glucagon Secretion 9
Somatostatin—Its Effect on Inhibition of Glucagon and Insulin Secretion 9
Pancreatic Polypeptide 9
Summary of Blood Glucose Regulation 9
Diabetes Mellitus 9
Pathological Physiology of Diabetes Mellitus 9
Physiology of Diagnosis 9
Treatment of Diabetes 9
Diabetic Coma 9
Hyperinsulinism 9

79 PARATHYROID HORMONE, CALCITONIN, CALCIUM AND PHOSPHATE METABOLISM, VITAMIN D, BONE, AND TEETH 9
Calcium and Phosphate in the Extracellular Fluid and Plasma—Function of Vitamin D 9
Absorption and Excretion of Calcium and Phosphate 9
Vitamin D and Its Role in Calcium and Phosphate Absorption 9
The Calcium in the Plasma and Interstitial Fluid 9
The Inorganic Phosphate in the Extracellular Fluids 9

Effects of Altered Calcium and Phosphate Concentrations in the Body Fluids 940
Bone and Its Relationship to Extracellular Calcium and Phosphates 941
Precipitation and Absorption of Calcium and Phosphate in Bone—Equilibrium with the Extracellular Fluids 941
Exchangeable Calcium 942
Deposition and Absorption of Bone—Remodeling of Bone 942
Parathyroid Hormone 943
Effect of Parathyroid Hormone on Calcium and Phosphate Concentrations in the Extracellular Fluid 944
Control of Parathyroid Secretion by the Calcium Ion Concentration 946
Calcitonin 946
Overall Control of Calcium Ion Concentration 948
Physiology of Parathyroid and Bone Diseases 948
Hypoparathyroidism 948
Hyperparathyroidism 949
Rickets 950
Osteoporosis 950
Physiology of the Teeth 950
Function of tha Different Parts of the Teeth 951
Dentition 951
Mineral Exchange in Teeth 952
Dental Abnormalities 952

80 REPRODUCTIVE AND HORMONAL FUNCTIONS OF THE MALE 954
Spermatogenesis 954
The Steps of Spermatogenesis 954
Function of the Seminal Vesicles 957
Function of the Prostate Gland 957
Semen 957
Abnormal Spermatogenesis and Male Fertility 957
The Male Sexual Act 958
Neuronal Stimulus for Performance of the Male Sexual Act 958
Stages of the Male Sexual Act 959
Testosterone and Other Male Sex Hormones 960
Secretion, Metabolism, and Chemistry of the Male Sex Hormone 960
Functions of Testosterone 961
Basic Intracellular Mechanism of Action of Testosterone 963
Control of Male Sexual Functions by the Gonadotropic Hormones—FSH and LH 963
Abnormalities of Male Sexual Function ... 964
The Prostate Gland and Its Abnormalities 964
Hypogonadism in the Male 965
Testicular Tumors and Hypergonadism in the Male 965
The Pineal Gland—Its Function in Controlling Seasonal Fertility 966

81 FEMALE PHYSIOLOGY BEFORE PREGNANCY, AND THE FEMALE HORMONES 968
Physiologic Anatomy of the Female Sexual Organs 968
The Female Hormonal System 968
The Monthly Ovarian Cycle and Function of the Gonadotropic Hormones 969
The Corpus Luteum—The "Luteal" Phase of the Ovarian Cycle 971
Functions of the Ovarian Hormones—Estradiol and Progesterone ... 972
Chemistry of the Sex Hormones 972
Functions of the Estrogens—Effects on the Primary and Secondary Sexual Characteristics 974
Functions of Progesterone 975
The Endometrial Cycle and Menstruation 975
Regulation of the Female Monthly Rhythm—Interplay Between the Ovarian and Hypothalamic-Pituitary Hormones ... 977
Feedback Oscillation of tha Hypothalamic-Pituitary-Ovarian System 978
Puberty and Menarche 978
The Menopause 979
Abnormalities of Secretion by the Ovaries 979
The Female Sexual Act 980
Female Fertility 981

82 PREGNANCY AND LACTATION 983
Maturation of the Ovum 983
Transport, Fertilization, and Implantation of the Developing Ovum 983
Early Intrauterine Nutrition of the Embryo 985
Function of the Placenta 985
Developmental and Physiologic Anatomy of the Placenta 985
Permeability of the Placental Membrane 986
Hormonal Factors in Pregnancy 987
Human Chorionic Gonadotropin and Its Effect in Causing Persistence of the Corpus Luteum and in Preventing Menstruation 987
Secretion of Estrogens by the Placenta 988
Secretion of Progesterone by the Placenta 988
Human Chorionic Somatomammotropin 989
Other Hormonal Factors in Pregnancy 989

Response of the Mother's Body to Pregnancy 990
Preeclampsia and Eclampsia 991
Parturition 991
Increased Uterine Excitability Near Term 991
Onset of Labor—A Positive Feedback Theory for Its Initiation 992
Abdominal Muscle Contractions During Labor 993
Mechanisms of Parturition 993
Separation and Delivery of the Placenta 993
Labor Pains 993
Involution of the Uterus 994
Lactation 994
Development of the Breasts 994
Initiation of Lactation—Function of Prolactin 994
The Ejection or "Let-Down" Process in Milk Secretion—Function of Oxytocin 995
Milk and the Metabolic Drain on the Mother Caused by Lactation 995

83 SPECIAL FEATURES OF FETAL AND NEONATAL PHYSIOLOGY 99
Growth and Functional Development of the Fetus 99
Development of the Organ Systems 99
Adjustments of the Infant to Extrauterine Life 99
Onset of Breathing 99
Circulatory Readjustments at Birth 100
Nutrition of the Newborn Infant 100
Special Functional Problems in the Neonatal Infant 100
Special Problems of Prematurity 100
Immature Development of the Premature Infant 100
Instability of the Control Systems in the Premature Infant 100
Danger of Blindness Caused by Oxygen Therapy in the Premature Infant 100
Growth and Development of the Child 100
Behavioral Growth 100

XIV SPORTS PHYSIOLOGY 100

84 SPORTS PHYSIOLOGY 1008
The Muscles in Exercise 1009
Strength, Power, and Endurance of Muscles 1009
The Muscle Metabolic Systems in Exercise 1009
Respiration in Exercise 101
The Cardiovascular System in Exercise 101
Body Heat in Exercise 101
Body Fluids and Salt in Exercise 101
Drugs and Athletes 101

INDEX 1019

I

INTRODUCTION TO PHYSIOLOGY: THE CELL AND GENERAL PHYSIOLOGY

1 ■ Functional Organization of the Human Body and Control of the "Internal Environment"

2 ■ The Cell and Its Function

3 ■ Genetic Control of Protein Synthesis, Cell Function, and Cell Reproduction

1

Functional Organization of the Human Body and Control of the "Internal Environment"

Physiology is the study of function in living matter Especially, it attempts to explain the physical and chemical factors that are responsible for the origin, development, and progression of life. Each type of life, from the very simple virus up to the largest tree or to the complicated human being, has its own functional characteristics. Therefore, the vast field of physiology can be divided into *viral physiology, bacterial physiology, cellular physiology, plant physiology, human physiology,* and many more subdivisions

Human Physiology. In human physiology we attempt to explain the specific characteristics and mechanisms of the human body that make it a living being The very fact that we remain alive is almost beyond our own control, for hunger makes us seek food and fear makes us seek refuge. Sensations of cold make us provide warmth, and other forces cause us to seek fellowship and to reproduce Thus, the human being is actually an automaton, and the fact that we are sensing, feeling, and knowledgeable beings is part of this au- - sequence of life; these special attributes . us to exist under widely varying conditions otherwise would make life impossible.

CELLS AS THE LIVING UNITS OF THE BODY

The basic living unit of the body is the cell, and each organ is an aggregate of many different cells held together by intercellular supporting structures Each type of cell is specially adapted to perform one particular function. For instance, the red blood cells, 25 trillion in all, transport oxyg from the lungs to the tissues. Though this type cell is perhaps the most abundant, there are a proximately another 50 trillion cells. The enti body, then, contains about 75 trillion cells

Though the many cells of the body often diff markedly from each other, all of them have certai basic characteristics that are alike For instanc in all cells oxygen combines with carbohydrat fat, or protein to release the energy required f cell function. Furthermore, the general mech nisms for changing nutrients into energy are ba ically the same in all cells, and all the cells als deliver the end-products of their chemical rea tions into the surrounding fluids.

Almost all cells also have the ability to repro duce, and whenever cells of a particular type ar destroyed from one cause or another, the remain ing cells of this type usually divide again an again until the appropriate number is replenished

THE EXTRACELLULAR FLUID— THE INTERNAL ENVIRONMENT

About 56 per cent of the adult human body is fluid. Though most of this fluid is inside the cells and is called *intracellular fluid*, about one third is in the spaces outside the cells and is called *extracellular fluid*. This extracellular fluid is in constant motion throughout the body. It is rapidly transported throughout the body by the blood circulation and then mixed between the blood and the tissue fluids by diffusion through the capillary walls. In the extracellular fluid are the ions and

nutrients needed by the cells for maintenance of cellular life. Therefore, all cells live in essentially the same environment, that is, the extrocellular fluid, for which reason the extracellular fluid is called the *internal environment* of the body, or the *milieu intérieur*, a term introduced a hundred years ago by the great 19th century French physiologist Claude Bernard.

Cells are capable of living, growing, and providing their special functions so long as the proper concentrations of oxygen, glucose, the different ions, amino acids, and fatty substances are available in this internal environment.

Differences Between Extrncellular and Intrncellular Fluide. The extracellular fluid contains large amounts of sodium, chloride, nnd bicarbonate ions, plus nutrients for the cells, such as oxygen, glucose, fatty acids, and amino acids. It also contains carbon dioxide that is being transported from the cells to the lungs to be excreted and other cellular products that are being transported to the kidneys for excretion.

The intracellular fluid differs significantly from the extracellular fluid; particularly, it contains large amounts of potassium, mogrnesium, and phosphate ions instead of the sodium and chloride ions found in the extracellular fluid. Special mechanisms for transporting ions through the cell membranes maintain these differences. These mechanieme will be discussed in Chapter 9.

"HOMEOSTATIC" MECHANISMS OF THE MAJOR FUNCTIONAL SYSTEMS

HOMEOSTASIS

The term *homeostasis* is used by physiologists to mean *maintenance of stotic*, or *constont, conditions in the internal environment*. Essentially all the organs and tissues of the body perform functions that help to maintain these constant conditions. For instance, the lungs provide oxygen to the extracellular fluid to replenish continually the oxygen that is being used by the cells, the kidneys maintain constant ion concentrations, and the gastrointestinal system provides nutrients. A large segment of this text is concerned with the manner in which each organ or tissue contributes to homeostasis. To begin this discussion, the different functional systems of the body and their homeostatic mechanisms will be outlined briefly; then the basic theory of the control systems that cause the functional systems to operate in harmony with each other will be discussed.

THE EXTRACELLULAR FLUID TRANSPORT SYSTEM—THE CIRCULATORY SYSTEM

Extracellular fluid is transported through all parts of the body in two different stages. The first stage entails movement of blood around and around the circulatory system, and the second, movemeat of fluid between the blood capillaries and the cells. Figure 1–1 illustrates the overall circulation of blood, showing that the heart is actually two separate pumps, one of which propels blood through the lungs and the other through the systemic circulation. All the blood in the circulation traverses the entire circuit of the circulation an average of once each minute at rest and as many as six times each minute when a person becomes extremely active.

As blood passes through the capillaries, continual fluid exchange occurs between the plasma portion of the blood and the interstitial fluid that fills the spaces between the cells, the intercellular spaces. This process is illustrated in Figure 1–2. Note that the capillaries are porous so that large amounts of fluid and its dissolved constituents can *diffuse* back and forth between the blood and the tissue spaces, as illustrated by the arrows. This process of diffusion is caused by kinetic motion of the molecules in both the plasma and the interstitial fluid. That is, the fluid and dissolved molecules are continuolly moving and bouncing in all directions within the fluid itself and olso through the pores and through the tissue spaces. Almost no cell is located more than 25 to 50 microns from a capillary, which insuree diffusion of almoet any substance from the capillary to the cell within a few seconds. Thus, the extracellular fluid everywhere in the body, both that of the plasma aad that in interstitial spaces, ie contiaually being

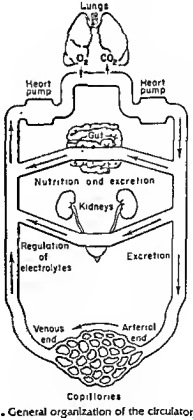


Figure 1–1. General organization of the circulatory system.

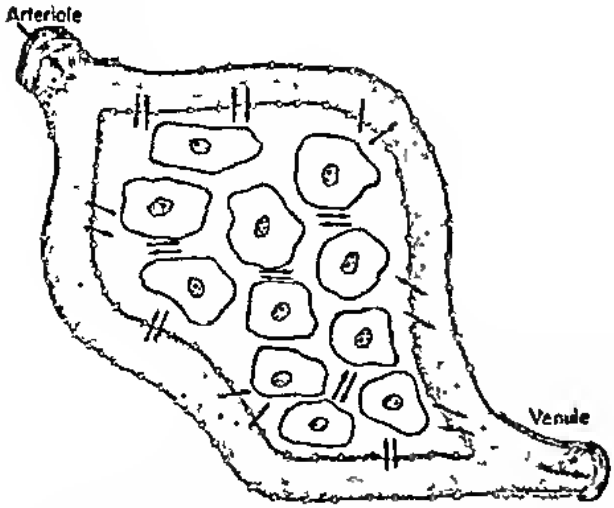


Figure 1–2. Diffusion of fluids through the capillary walls and through the interstitial spaces

mixed, thereby maintaining almost complete homogeneity.

ORIGIN OF NUTRIENTS IN THE EXTRACELLULAR FLUID

The Respiratory System. Figure 1–1 shows that each time the blood passes through the body it also flows through the lungs. The blood picks up oxygen in the alveoli, thus acquiring the oxygen needed by the cells. The membrane between the alveoli and the lumen of the pulmonary capillaries is only 0.4 to 2.0 microns in thickness, and oxygen diffuses through this membrane into the blood in exactly the same manner that water and the ions diffuse through the tissue capillaries

The Gastrointestinal Tract. A large portion of the blood pumped by the heart passes also through the walls of the gastrointestinal organs. Here different dissolved nutrients, including carbohydrates, fatty acids, amino acids, and others, are absorbed into the extracellular fluid.

The Liver and Other Organs That Perform Primarily Metabolic Functions. Not all substances absorbed from the gastrointestinal tract [illegible] be used in their absorbed form by the cells. [illegible] liver changes the chemical compositions of [illegible] of these to more usable forms, and other [illegible]ssues of the body—the fat cells, the gastrointestinal mucosa, the kidneys, and the endocrine glands—help to modify the absorbed substances or store them until they are needed at a later time.

Musculoskeletal System. Sometimes the question is asked. How does the musculoskeletal system fit into the homeostatic functions of the body? The answer to this is obvious and simple: Were it not for this system, the body could not move to the appropriate place at the appropriate time to obtain the foods required for nutrition. The musculoskeletal system also provides motility for protection against adverse surroundings, without which the entire body, and along with it all the homeostatic mechanisms, could be destroyed instantaneously.

REMOVAL OF METABOLIC END-PRODUCTS

Removal of Carbon Dioxide by the Lungs. At the same time that blood picks up oxygen in the lungs, carbon dioxide is released from the blood into the alveoli, and the respiratory movement of air into and out of the alveoli carries the carbon dioxide to the atmosphere. Carbon dioxide is the most abundant of all the end-products of metabolism.

The Kidneys. Passage of the blood through the kidneys removes most substances from the plasma that are not needed by the cells. These substances include especially different end-products of cellular metabolism and excesses of ions and water that might have accumulated in the extracellular fluid. The kidneys perform their function by, first, filtering large quantities of plasma through the glomeruli into the tubules and then reabsorbing into the blood those substances needed by the body such as glucose, amino acids, appropriate amounts of water, and many of the ions. However, most of the substances not needed by the body, especially the metabolic end-products such as urea, are not reabsorbed but, instead, pass on through the renal tubules into the urine.

REGULATION OF BODY FUNCTIONS

The Nervous System. The nervous system is composed of three major parts: the *sensory portion*, the *central nervous system* (or *integrative portion*) and the *motor portion*. Sensory receptors detect the state of the body or the state of the surroundings For instance, receptors present everywhere in the skin apprise one every time an object touches that person at any point. The eyes are sensory organs that give one a visual image of the surrounding area. The ears also are sensory organs. The central nervous system is composed of the brain and spinal cord. The brain can store information, generate thoughts, create ambition, and determine reactions the body performs in response to the sensations. Appropriate signals are then transmitted through the motor portion of the nervous system to carry out the person's desires.

A large segment of the nervous system is called the *autonomic system*. It operates at a subconscious level and controls many functions of the internal organs, including the action of the heart, the movements of the gastrointestinal tract, and the secretion by different glands.

The Hormonal System of Regulation. Located in the body are eight major endocrine glands that secrete chemical substances, the *hormones*. Hormones are transported in the extracellular fluid to

l parts of the body to help regulate cellular nction. For instance, thyroid hormone increases e rates of most chemical reactions in all cells. In is way thyroid hormone helps to set the tempo bodily activity. Likewise, insulin controls glu- se metabolism; adrenocortical hormones control n and protein metabolism; and parathormone ntrols bone metabolism. Thus, the hormones are system of regulation that complements the ner- ıus system. The nervous system, in general, reg- ıates mainly muscular and secretory activities of ıe body, whereas the hormonal system regulates ainly the metabolic functions.

EPRODUCTION

Reproduction sometimes is not considered to be bomeostatic function. But it does help to main- ıin static conditions by generating new beings to ıke the place of ones that are dying. This perhaps ɔunds like a farfetched usage of the term homeo- tasis, but it does illustrate that, in the final nalysis, essentially all structures of the body are ɔ organized that they help maintain the automa- icity and the continuity of life.

HE CONTROL SYSTEMS ɔF THE BODY

The buman body has literally thousands of con- ɔl systems in it. The most intricate of all these re the genetic control systems that operate within ll cells to control intracellular function, and also ɔ control all life processes, a subject that will be iscussed in detail in Chapter 3. Many other con- rol systems operate within the organs to control unctions of the individual parts of the organs; thers operate throughout the entire body to con- rol the interrelationships between the organs. For nstance, the respiratory system, operating in as- ociation with the nervous system, regulates the oncentration of carbon dioxide in the extracellular luid. The liver and the pancreas regulate the oncentration of glucose in the extracellular fluid. And the kidneys regulate the concentrations of ıydrogen, sodium, potassium, phosphate, and ɔther ions in the extracellular fluid.

XAMPLES OF CONTROL MECHANISMS

Regulation of Oxygen and Carbon Dioxide Concentrations in the Extracellular Fluid. Since oxygen is one of the major substances re- quired for chemical reactions in the cells, it is ortunate that the body has a special control mech- anism to maintain an almost exact and constant oxygen concentration in the extracellular fluid. This mechanism depends principally on the chem- ical characteristics of *hemoglobin*, which is present in all the red blood cells. Hemoglobin combines with oxygen as the blood passes through the lungs. Then, as the blood passes through the tissue capillaries the hemoglobin will not release oxygen into the tissue fluid if too much oxygen is already there, but if the oxygen concentration is too little, sufficient oxygen will be released to re-establisb an adequate tissue oxygen concentration. Thus, the regulation of oxygen concentration in the tissues is vested principally in the chemical characteristics of hemoglobin itself. This regulation is called the oxygen-*buffering function of hemoglobin*.

Carbon dioxide concentration in the extracellular fluid is regulated in quite a different way. Carbon dioxide is one of the major end-products of the oxidative reactions in cells. If all the carbon dioxide formed in the cells should continue to accumulate in the tissue fluids, the mass action of the carbon dioxide itself would soon halt all the energy-giving reactions of the cells. Fortunately, a nervous mechanism controls tbe expiration of carbon dioxide through the lungs and in this way maintains a constant and reasonable concentration of carbon dioxide in the extracellular fluid. That is, a high carbon dioxide concentration *excites the respiratory center*, causing the person to breathe rapidly and deeply. This increases the expiration of carbon dioxide and therefore increases its removal from the blood and extracellular fluid, and the process continues until the concentration returns to normal.

Regulation of Arterial Pressure. Several different systems contribute to the regulation of arterial pressure. One of these, the *baroreceptor system*, is a very simple and excellent example of a control mechanism. In the walls of most of the great arteries of the upper body, especially the bifurcation region of the carotids and the arch of the aorta, are many nerve receptors, called *baroreceptors*, which are stimulated by stretch of the arterial wall. When the arterial pressure becomes great, these baroreceptors are stimulated excessively, and impulses are transmitted to the medulla of the brain. Here the impulses inhibit the *vasomotor center*, which in turn decreases the number of impulses transmitted through the sympathetic nervous system to the heart and blood vessels. Lack of these impulses causes diminished pumping activity by the heart and increased ease of blood flow through the peripheral vessels, both of which lower the arterial pressure back toward normal. Conversely, a fall in arterial pressure relaxes the stretch receptors, allowing the vasomotor center to become more active than usual and thereby causing the arterial pressure to rise back toward normal.

CHARACTERISTICS OF CONTROL SYSTEMS

The above examples of homeostatic control mechanisms are only a few of the many hundreds in the body, all of which have certain characteristics in common. These are explained in the following pages.

Negative Feedback Nature of Control Systems. Most control systems of the body act by a process of *negative feedback*, which can be explained best by reviewing some of the homeostatic control systems mentioned earlier. In the regulation of carbon dioxide concentration, a high concentration of carbon dioxide in the extracellular fluid causes increased pulmonary ventilation, and this in turn causes decreased carbon dioxide concentration In other words, the high conrentration causes a decreased concentration, which is *negative* to the initiating stimulus Conversely, if the carbon dioxide concentration falls too low, this causes a feedback increase in the concentration. This response also is negative to the initiating stimulus.

In the arterial pressure-regulating mechanisms, a high pressure causes a series of reactions that promote a lowered pressure, or a low pressure causes a series of reactions that promote an elevated pressure In both instances these effects are negative with respect to the initiating stimulus

Therefore, in general, if some factor becomes excessive or too little, a control system initiates *negative feedback*, which consists of a series of changes that return the factor toward a rertain mean value, thus maintaining homeostasis.

"Gain" of a Control System. The degree of effectiveness with which a control system maintains constant conditions is determined by the *gain* of ths negative feedback. For instanre, let us assume that a large volume of blood is transfused into a psrson whoss baroreceptor pressure control system is not functioning, and the arterial pressure rises from the normal level of 100 mm Hg up to 175 mm Hg. Then, assume the same volume of blood is injected into the same person when his baroreceptor system is functioning and this time the pressure rises only 25 mm Hg Thus, the feedback control system has caused a "correction" of −50 mm Hg—that is, from 175 mm Hg to 125 mm Hg. Yet, there still remains an increase in pressure of +25 mm Hg, called the "error," which means that the control system is not 100 per cent effective in preventing change The gain of the system is then calculated by the following formula:

$$\text{Gain} = \frac{\text{Correction}}{\text{Error}}$$

Thus, in the above example, the correction is −50 mm Hg, and the error still persisting is +25 mm Hg. Therefore, the gain of the person's baroreceptor system for control of arterial pressure is −50 divided by +25, or −2. That is, an extraneous factor that tends to increase or decrease the arterial pressure does so only one third as much as would occur if this control system were not present.

The gains of some other physiological contcol systems are much greater than that of the baroreceptor system For instance, the gain of the system controlling body temperature is approximately −33. Therefore, one can see that the temperature control system is much more effecti than the baroreceptor pressure control system

Positive Feedback—Vicious Circles and Death Caused by Positive Feedback. One might ask the question: why do essentially al control systems of the body operate by negativ feedback rather than positive feedback? However if you will consider the nature of positive feedback you will immediately see that positive feedbad does not lead to stability but to instability an often to death

Figure 1–3 illustrates an instance in whic death can ensue from positive feedback. This figur depicts the pumping effectiveness of the heart ahowing that the heart of the normal human bein pumps about 5 liters of blood per minute. However if the person is suddenly bled 2 liters, the amoun of blood in the body is decreased to such a lo level that not enough is available for the heart t pump effectively. As a result, the arterial pressur falls, and the flow of blood to the heart musc through the coronary vessels also diminishes. Thi results in weakening of the heart, further dimin ished pumping, further decrease in coronary bloo flow, and still more weakness of the heart; th cycle repeats itself again and again until deat Note that each cycle in the feedback results i further weakening of the heart. In other word the initiating stimulus causes more of the same which is *positive feedback*.

Positive feedback is better known as a "vicio circle," but actually a mild degree of positive feed back can be overcome by the negative feedbac control mechanisms of the body, and a vicio circle will fail to develop. For instance, if th person in the above example were bled only 1 lite instead of 2 liters, the normal negative feedbac mechanisms for controlling cardiac output an arterial pressure would overbalance the positiv feedback, and the person would recover, as show by the dashed curve of Figure 1–3.

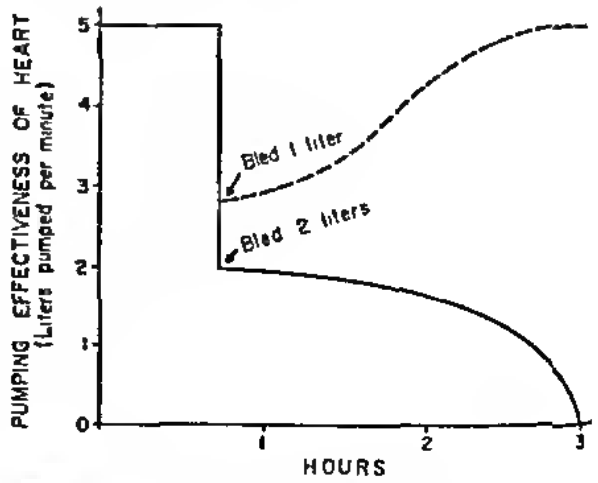


Figure 1–3. Death caused by positive feedback when 2 liters of blood are removed from the circulation

OF THE BODY

The purpose of this chapter has been to point it, first, the overall organization of the body and, cond, the means by which the different parts of e body operate in harmony. To summarize, the ody is actually a *social order of about 75 trillion lls* organized into different functional structures, me of which are called *organs*. Each functional ructure provides its share in the maintenance of omeostatic conditions in the extracellular fluid, hich is called the *internal environment*. As long s normal conditions are maintained in the inter- al environment, the cells of the body will continue live and function properly. Thus, each cell ben- its from homeostasis, and in turn each cell con- ributes its share toward the maintenance of omeostasis. This reciprocal interplay provides ontinuous automaticity of the body until one or ore functional systems lose their ability to con- ribute their share of function. When this happens, ll the cells of the body suffer. Extreme dysfunction eads to death, while moderate dysfunction leads o sickness.

PPENDIX

ASIC PHYSICAL PRINCIPLES OF CONTROL SYSTEMS

Until recently the principles of control systems as pplied to the human body have been taught only ualitatively rather than quantitatively. For instance, n the case of baroreceptor regulation of arterial pres- ure, physiologists have simply stated that an increase n pressure causes a reflex decrease in pressure toward ormal. However, this means very little unless we know *ow much* effect occurs But within the past 30 years, pplication of much more quantitative physical princi- les of control systems has begun to make physiology a nuch more exact science than it has been in the past.

Basic Symbols Used in Control System Analysis. igure 1–4 illustrates common basic symbols used in ontrol system analysis. They are the following:

The Addition-Subtraction Symbol. This symbol is hown in Figure 1–4A. For instance, let x represent the ate of intake of salt in the solid food eaten each day, y he intake of salt in liquids drunk each day, and z the ate of loss of salt in the urine each day. Then the net ate of change of salt in the body will be x + y − z, as ndicated by the arrow.

The Multiplication Symbol. Figure 1–4B illustrates multiplication of three quantities, x, y, and z, to give xyz. For instance, let us assume that three separate factors are affecting arterial pressure and that these factors multiply each other. Thus, the baroreceptor system might be causing x effect to elevate arterial pressure; a hormone secreted by one of the endocrine glands might be causing y effect to elevate arterial pressure; and hemorrhage might be causing z effect to decrease arterial pressure. The net effect would be xyz.

Multiplication by a Constant Factor. Figure 1–4C illustrates multiplication by a constant factor. Let us assume that x is the *concentration* of sodium in the extracellular fluids and K is the volume of extracellular

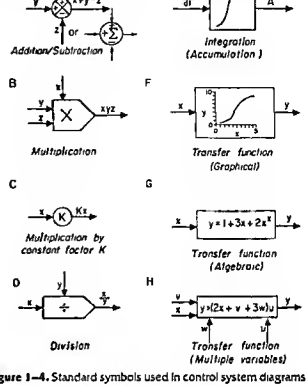


Figure 1–4. Standard symbols used in control system diagrams

fluid; the total *quantity* of sodium in the extracellular fluid would then be Kx

The Division Symbol. Figure 1–4D shows the value x divided by the value y to give x/y. For example, if x is the total *quantity* of sodium in the extracellular fluid and y is the extracellular fluid *volume*, then the output of this block, x/y, is the concentration of sodium in the extracellular fluid.

The Integration (or Accumulation) Symbol. Many functions of the body depend upon slow accumulation of some factor Thus, in Figure 1–4E, if the *rate of change* of aldosterone in the body is represented by the differential term, dA/dt, the output of the block is the *quantity* of aldosterone that has accumulated in the body at any given time, t. The symbol in this block is called the *integration symbol*, or, in other words, the rate dA/dt is integrated to give A. If dA/dt is positive, the quantity of aldosterone in the body will be increasing, whereas if dA/dt is negative, the quantity of aldosterone in the body will be decreasing.

Other Types of Mathematical Functions. Figures 1–4F, G, and H illustrate other types of mathematical functions in which x is shown in each example to enter a block and y to exit from the block. Each block means that y is related to x in accordance with the function inside the box In Figure 1–4F this function is represented graphically. In Figure 1–4G it is represented by an algebraic equation. Figure 1–4H illustrates four inputs and one output, showing that three of the inputs add to each other and the other multiplies the first three.

As an example, Figure 1–4F might represent the relationship between glucose concentration, x, in the extracellular fluids and the rate of insulin secretion, y, by the pancreas. This function shows that at low glucose concentrations essentially no insulin is secreted, but at

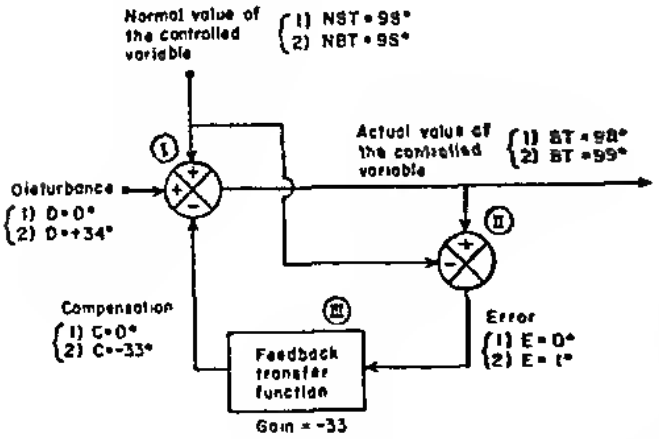


Figure 1–5. Static analysis of a control system. This figure presenting an analysis of the control of body temperature, showing values for different variables in the system under two different conditions (1) normal, and (2) when there is an increase in temperature of 34° F, thereby creating a disturbance that attempts to raise the body temperature (NBT, normal body temperature; actual body temperature, E, error, C, compensation, D, disturbance

high concentrations very large quantities of insulin are secreted

General Analysis of a Control System. Figure 1–5 illustrates a general analysis that can be applied to almost any negative feedback control system in the body The goal of most control systems is to keep some *controlled variable* at an almost constant value In Figure 1–5 the controlled variable is the body temperature (BT), and this is represented by the quantitative values on the arrow projecting to the right The three mathematical components of the control system are the following:

Block I illustrates the summation of the three different factors that determine the value of the controlled variable The three inputs to this block are (1) the *normal value for the controlled variable*, that is, the mean level at which the control system attempts to maintain the variable; (2) any *disturbance* that is acting on the body to cause the variable to deviate from its normal value, and (3) any *compensation* that is caused by the control system to counteract the disturbance.

Block II calculates the difference between the actual value of the controlled variable and the normal value The output of this block is called the *error*

Block III is the feedback portion of the control system that responds to the error and determines the degree of compensation that will occur to overcome the disturbance

Application of the General Analysis of Control of Body Temperature. Now let us apply this general analysis of a control system to an actual example, the control of body temperature At each point in the system two different sets of values are shown The first value in each instance represents the normal when there is no disturbance and no compensation by the control system Thus, the normal values are as follows

Normal value for the controlled variable ..	98° (F)
Disturbance . .	0°
Actual value of the controlled variable	98°
Error .	0°
Compensation	0°

Next, let us add a disturbance that attempts to change the body temperature In this instance, we will assume that the air temperature increases by 34° If the body temperature control system were not functioning, this would increase the body temperature also by 34°. However, as the body temperature begins to rise, this becomes greater than the normal value of 98°, and error develops Then the error activates the system, which in turn provides enough compensation overcome most of the disturbing effect of the air temperature. This compensation results from creased blood flow to the skin, decreased rate of production in the body, increased evaporation of swe and several other effects, all of which tend to reduce body temperature back toward normal Therefore, second set of values shown in Figure 1–5 is the following

Normal value of the controlled variable ...	98°
Disturbance	+34°
Actual value of the controlled variable ...	99°
Error	1°
Compensation	33°

Thus, one sees that after the control system become activated, the actual value of the control variable becomes equal to the sum of three different values (1) the normal value for the controlled variable (98°), (2) the disturbance (+34°), and (3) the compensation (−33°), giving an actual value of the control variable of 99° That is, the abnormal disturbance +34° causes a rise in body temperature of only 1°.

One can then calculate the gain of this system by dividing the compensation by the error (−33 divided by 1°), yielding a feedback gain of −33)

Figure 1–6 illustrates the effect of this 34° increase

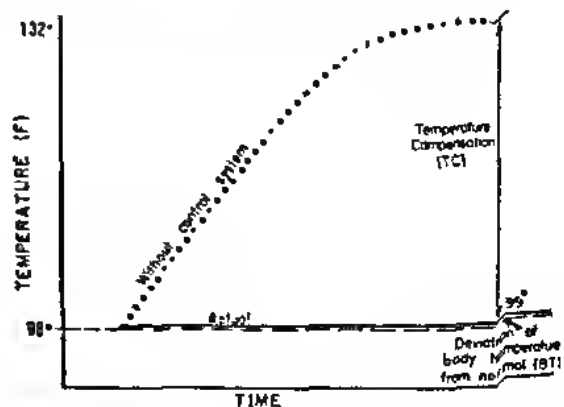


Figure 1–6. Effects on body temperature of suddenly increasing the air temperature 34° F, showing the hypothetical effect with a control system and the actual effect with a normal control system.

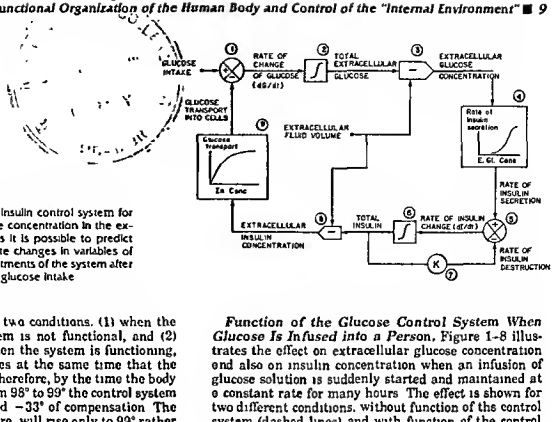


Figure 1–7. Analysis of the insulin control system for maintaining a constant glucose concentration in the extracellular fluid. By this analysis it is possible to predict transient as well as steady-state changes in variables of the system, such as the readjustments of the system after sudden changes in the rate of glucose intake

in air temperature under two conditions. (1) when the control system is not functional, and (2) when it is functional. When the system is functioning, the compensation increases at the same time that the body temperature rises. Therefore, by the time the body temperature has risen from 98° to 99° the control system will have already initiated −33° of compensation. The body temperature, therefore, will rise only to 99° rather than to 132°, an increase that would occur if the control system did not exist.

More Complex Analysis of a Control System—The Glucose Control System. Figure 1–7 illustrates an analysis of the insulin control system for regulating glucose concentration in the extracellular fluid. The basic system is the following: When a person eats increased quantities of glucose, the rising glucose concentration in the extracellular fluid causes the pancreas to secrete increased quantities of insulin. The insulin, in turn, causes increased transport of glucose through the cell membranes to the interior of the cells in most parts of the body. The glucose is then used for energy. This, obviously, returns the extracellular glucose concentration back toward normal. A quantitative mathematical analysis is illustrated by the nine blocks of Figure 1–7; these blocks can be explained as follows:

Block 1 calculates the rate of change of glucose (dG/dt) in the extracellular fluids by subtracting rate of glucose transport into the cells from the rate of intake of glucose.

Block 2 integrates the rate of change of glucose to give the total extracellular glucose at any given time.

Block 3 calculates the extracellular glucose concentration by dividing the total extracellular glucose by the extracellular fluid volume.

Block 4 illustrates the effect of extracellular glucose concentration on rate of insulin secretion

Block 5 sums the rate of insulin secretion and the rate of insulin destruction to give the rate of insulin change (dI/dt).

Block 6 integrates the rate of insulin change to give the total insulin in the body at any given time.

Block 7 calculates rate of insulin destruction by multiplying the total insulin by the constant K.

Block 8 calculates the extracellular insulin concentration by dividing total insulin by extracellular fluid volume.

Block 9 gives the effect of extracellular insulin at different concentrations on the rate of glucose transport into the body cells.

Function of the Glucose Control System When Glucose Is Infused into a Person. Figure 1–8 illustrates the effect on extracellular glucose concentration and also on insulin concentration when an infusion of glucose solution is suddenly started and maintained at a constant rate for many hours. The effect is shown for two different conditions, without function of the control system (dashed lines) and with function of the control system (solid line). Note that without the control system the glucose concentration rises to approximately 170 mg per cent and stabilizes at this level. Note, too, that there is also no increase in secretion of insulin.

On the other hand, when the control system is operative the glucose concentration begins to rise just as rapidly as before, but it soon stops rising and returns to a value not far from the normal glucose concentration level. This is caused by the effect of the glucose on the pancreas to cause insulin secretion, followed by buildup of insulin in the body fluids, until glucose transport through the cell membranes rises to equal the rate of glucose infusion into the blood. Note that the insulin concentration remains elevated as long as glucose continues to be infused, a condition that is necessary if the glucose concentration in the body fluids is to remain near normal. Note also that at first the glucose concentration slightly overshoots the final level at which it stabilizes. This is caused by the initial buildup of glucose in the extracellular fluid before the insulin has had time to become fully active

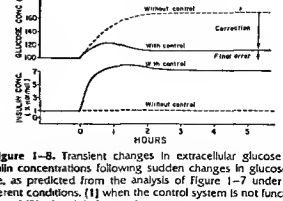


Figure 1–8. Transient changes in extracellular glucose and insulin concentrations following sudden changes in glucose intake, as predicted from the analysis of Figure 1–7 under two different conditions, (1) when the control system is not functioning, and (2) when it is functioning.

To the right in Figure 1–8 the degree of correction caused by the glucose control system is illustrated by the downward-directed arrow. The final error is illustrated by the upward-directed arrow The ratio of the lengths of these two arrows represents the gain of the system Since the correction arrow is approximately six times as long as the error arrow, the gain is approximately −6.

Steady-State Versus Transient Analyses of Control Systems. The analysis of the system for body temperature regulation shown in Figure 1–5 was a steady-state analysis. That is, it simply showed the initial conditions and the final conditions but did not characterize the transient events occurring in the control system during its activation On the other hand, the analysis of glucose control in Figures 1–7 and 1–8 gives the transients through which the different elements of the system pass in developing the final steady-state condition In other words, the control system does not act instantly, but instead requires a certain amount of time to develop its compensation. Furthermore, the controlled variable often overshoots the final steady-state value before it stabilizes

REFERENCES

Adolph, E F Origins of Physiological Regulations New York, Academic Press, 1968

Adolph, E F Physiological adaptations Hypertrophies and super functions Am Sci 60 608, 1972

Adolph, E F Physiological integrations in action *The Physiologist*, *25* (Suppl) 1 1982

Bernard C Lectures on the Phenomena of Life Common to Animals and Plants Springfield, Ill , Charles C Thomas 1974

Borow, M Fundamentals of Homeostasis A Clinical Approach to Fluid Electrolyte Acid Base Energy Metabolism in Health and Disease Flushing N Y Medical Examination Publishing Co 1977

Brown, J H U (ed) Engineering Principles in Physiology Vols 1 and 2 New York, Academic Press, 1973

Bruni, C *et al* (eds) Systems Theory in Immunology New York Springer-Verlag 1979

Bryant, P J and Simpson, P Intrinsic and extrinsic control of growth in developing organs *Q Rev Biol*, *59* 387 1984

Buruttini, R, and Borgdorff, P Closed loop baroreflex control of total peripheral resistance in the cat. Identification of gains by aid of a model *Cardiovasc Res* 18 715, 1984

Cannon, W B The Wisdom of the Body New York, W. W. Norto Co , 1932

Dawes, I W , and Sutherland, I W Microbial Physiology Oxf Blackwell Scientific Publishers, 1976

Frisancho, A R Human Adaptation St Louis, C V. Mosby Co , 19

Gann, D S *et al* Neural interaction in control of adrenocorticotro *Fed Proc*, *44* 161, 1985

Guyton, A. C Arterial Pressure and Hypertension Philadelphia W Saunders Co , 1980

Guyton, A C , and Coleman, T G Quantitative analysis of the pa physiology of hypertension *Circ Res*, *14* I-1, 1969

Guyton, A. C , *et al* Cardiac Output and Its Regulation Philadelp W B Saunders Co , 1973

Guyton, A C , *et al* Dynamics and Control of the Body Fluids. Ph delphia, W B Saunders Co , 1975

Hemker, H C , and Hess, B , Analysis and Simulation of Biochem Systems New York, American Elsevier Publishing Co , 1972.

Huffaker, C B (ed) Biological Control New York, Plenum Press, 19

Iberall A S , and Guyton, A C (eds) *Proc Int Symp on Dynan and Controls in Physiological Systems* Regulation and Control Physiological Systems. ISA, Pittsburgh, 1973

Jones, R W Principles of Biological Regulation An Introduction Feedback Systems New York Academic Press, 1973

Klevecz, R B *et al* Cellular clocks and oscillators *Int Rev Cy* *86* 97, 1984

Krieger, D T , and Aschoff, J Endocrine and other biological rhyth *In* DeGroot, L H , *et al* (eds) Endocrinology Vol 3 New Y Grune & Stratton 1979, p 2079

Marmarelis, P Z and Marmarelis, V Z Analysis of Physiolog Systems The White-Noise Approach New York, Plenum Press, 19

McIntosh, J E A and McIntosh, R P Mathematical Modeling Computers in Endocrinology New York, Springer-Verlag 1980.

Milhorn, H T The Application of Control Theory to Physiolog Systems Philadelphia, W B Saunders Co , 1966

Miller, S L , and Orgel, L E The Origins of Life on the Ea Englewood Cliffs, N J , Prentice Hall, 1974

Piva, F *et al* Regulation of hypothalamic and pituitary function. Le short, and ultrashort feedback loops *In* DeGroot L J , *et al* (e Endocrinology Vol 1 New York Grune & Stratton, 1979, p 21

Randall, J E Microcomputers and Physiological Simulation Read Mass Addison Wesley Publishing Co 1980

Reeve, E B , and Guyton, A C Physical Bases of Circulatory Transp Regulation and Exchange Philadelphia W B Saunders Co , 196

Rusak, B , and Zucker, I Neural regulation of circadian rhyth *Physiol Rev*, *59* 449, 1979

Söderberg, U Neurophysiological aspects of homeostasis *Annu R Physiol* *26* 271, 1964

Sweetser, W Human Life (Aging and Old Age) New York, Arno Pr 1979

Toates, F M Control Theory in Biology and Experimental Psychol London, Hutchinson Education Ltd , 1975

Weston, L Body Rhythm The Circadian Rhythms Within You N York, Harcourt Brace Jovanovich, 1979

2

The Cell and Its Function

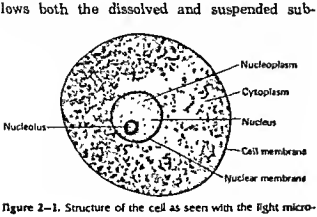
Each of the 75 trillion cells in the human being s a living structure that can survive indefinitely ınd, in most instances, can even reproduce itself, ırovided its surrounding fluids contain appropriate ıutrients. To understand the function of organs ınd other structures of the body, it is essential hat we first understand the basic organization of ha cell and the functions of its component parts.

ORGANIZATION OF THE CELL

A typical cell, as seen by tha light microscope, s illustrated in Figura 2–1. Its two major parts ıre the *nucleus* and the *cytoplasm*. The nucleus is ıeparated from tha cytoplasm by a *nuclear* mem-ɔrane, and tha cytoplasm is separated from the surrounding fluids by a *cell* membrane.

The different substances that make up the cell ara collectively called *protoplasm*. Protoplasm is composed mainly of five basic substances: water, electrolytes, proteins, lipids, and carbohydrates.

Water. The principal fluid medium of the cell is water, which is present in a concentration of between 70 and 85 per cent. Many cellular chemicals are dissolved in the water, while others are suspended in small particulate form. Chemical reactions take place among the dissolved chemicals or at the surface boundaries between the suspended particles and the water. The fluid nature of water sllows both the dissolved and suspended substances to diffuse or flow to different parts of the cell, thereby providing transport of the substances from one part of the cell to another.



ıgure 2–1. Structure of the cell as seen with the light micro-ıcope.

Electrolytes. The most important electrolytes in the cell are *potassium*, *magnesium*, *phosphate*, *sulfate*, *bicarbonate*, and small quantities of *sodium*, *chloride*, and *calcium*. These will be discussed in much greater detail in Chapter 9, which will consider the interrelationships between the intracellular and extracellular fluids.

The electrolytes are dissolved in the cell water, and they provide inorganic chemicals for cellular reactions. Also, they are necessary for operation of some of the cellular control mechanisms. For instance, electrolytes acting at the cell membrane allow transmission of electrochemical impulses in nerve and muscle fibers, and the intracellular electrolytes determine the activity of different enzymatically catalyzed reactions that ara necessary for cellular metabolism.

Proteins. Next to water, the most abundant substance in most cells is proteins, which normally constitute 10 to 20 per cent of the cell mass. Thesa can be divided into two different types, *structural proteins* and *globular proteins* that are mainly *enzymes*.

To get an idea of what is meant by *structurol proteins*, one needs only to note that leather is composed principally of structural proteins, and that hair is almost entirely a structural protein. Proteins of this type are present in the cell in the form of long thin filaments that themselves aro polymers of many protein molecules. The most prominent use of such intracellular filaments is to provide the contractile mechanism of all muscles. However, filaments are also organized into microtubules that provide the structures of such organelles as cilia and the mitotic spindles of mitosing cells. And, extracellularly, fibrillar proteins are found especially in the collagen and elastin fibers of connective tissue, blood vessels, tendons, ligaments, and so forth.

The *globular proteins*, on the other hand, are an entirely different type of protein, composed usually of individual protein molecules or at most aggre-

gates of a few molecules in a globular form rather than a fibrillar form. These proteins are mainly the enzymes of the cell and, in contrast to the fibrillar proteins, are often soluble in the fluid of the cell or are integral parts of or adherent to the membranous structures inside the cell. The enzymes come into direct contact with other substances inside the cell and catalyze chemical reactions. For instance, the chemical reactions that split glucose into its component parts and then combine these with oxygen to form carbon dioxide and water while at the same time providing energy for cellular function are catalyzed by a series of protein enzymes Thus, enzyme proteins control the metabolic functions of the cell.

Special types of proteins are present in different parts of the cell. Of particular importance are the *nucleoproteins*, present both in the nucleus and the cytoplasm The nucleoproteins of the nucleus contain *deoxyribonucleic acid (DNA)*, which constitutes the *genes*, and these control the overall function of the cell as well as the transmission of hereditary characteristics from cell to cell. These substances will be considered in detail in Chapter 3. In addition, the chemical nature of proteins will be considered in Chapter 69, and the different structural and enzymatic functions of proteins will be subjects of discussion at numerous pointa throughout this text.

Lipids. Lipids are several different types of substances that are grouped together because of their common property of being soluble in fat solvents. The most important lipids in most cells are *phospholipids* and *cholesterol*, which constitute about 2 per cent of the total cell mass. These are major constituents of the different membranes such as the cell membrane, the nuclear membrane, and the membranes lining intracytoplasmic organelles, e.g., the endoplasmic reticulum and the mitochondria The special importance of phospholipids and cholesterol in the cell is that they are either insoluble or only very slightly soluble in water

In addition to phospholipids and cholesterol, some cells contain large quantities of *triglycerides*, also called *neutral fat*. In the so-called "fat cells," triglycerides often account for as much as 95 per cent of the cell mass And this fat stored in these cells represents the body's main storehouse of energy-giving nutrient that can later be dissoluted and used for energy wherever in the body it is needed

The chemical natures of the different types of lipids and their functions in the body will be discussed in Chapter 68

Carbohydrates. In general, carbohydrates have very little structural function in the cell except as part of glycoprotein molecules, but they play a major role in nutrition of the cell. Most human cells do not maintain large stores of carbohydrates, usually averaging about 1 per cent of their total mass. However, carbohydrate, in the form of gl cose, is always present in the surrounding extr cellular fluid so that it is readily available to t cell. A small amount of carbohydrate is usual stored in the cells in the form of *glycogen*, whi is an insoluble polymer of glucose and can be us rapidly to supply the cells' energy needs.

PHYSICAL STRUCTURE OF THE CELL

The cell is not merely a bag of fluid, enzym and chemicals; it also contains highly organiz physical structures, many of which are called *ganelles*, and the physical nature of each of the is equally as important to the function of the c as the cell's chemical constituents. For instan without one of the organelles, the *mitochondr* more than 95 per cent of the energy supply of t cell would cease immediately. Some principal ganelles of tha cell are illustrated in Figure 2- including the *cell membrone*, *nuclear membra endoplasmic reticulum*, *mitochondria*, and *ly somes*. Others not shown in the figure are, t *Golgi apparatus*, *centrioles*, and *cilia*.

THE MEMBRANOUS STRUCTURES OF THE CELL

Essentially all physical structures of the cell a lined by membranes composed primarily of lipi and proteins. These membranes include the c *membrane*, the *nuclear membrane*, the *membra of the endoplasmic reticulum*, and the *membran of the mitochondria*, *lysosomes*, *Golgi apparat* as well as still others. The lipids of the membran provide a barrier that prevents free movement water and water-soluble substances from one c compartment to the other. The protein molecul in the membrane, on the other hand, often per trate all the way through the membrane, th interrupting the continuity of the lipid barrier a therefore provide pathways for passage of speci substances through the membrane. Also, many the membrane proteins are enzymes that catal a multitude of different chemical reactions wh will be the subjects of further discussions in t and future chapters.

The Cell Membrane — 7.5–10 nm thick

The cell membrane, which completely envel the cell, is a very thin, elastic structure only to 10 nanometers thick. It is composed alm entirely of proteins and lipids; the approxim composition is proteins, 55 per cent; phospholipi 25 per cent; cholesterol, 13 per cent; other lipi 4 per cent; and carbohydrates, 3 per cent.

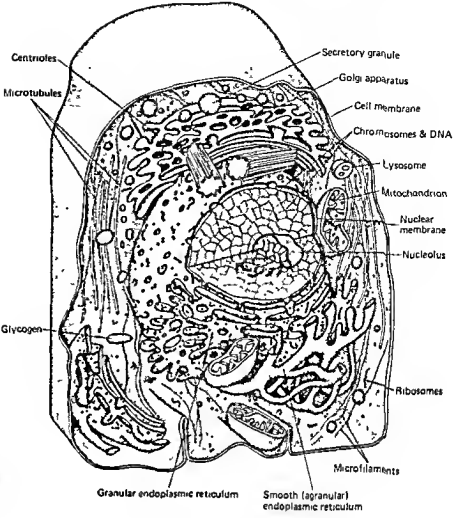


Figure 2–2. Reconstruction of a typical cell, ıowing the internal organelles in the cyto- lasm and in the nucleus.

The Lipid Barrier of the Cell Membrane. 'Igure 2–3 illustrates the cell membrane. Its basic tructure is a *lipid bilayer*, which is a thin film of ipids only 2 molecules thick that is continuous ıver the entire cell surface. Interspersed in this ipid film are large globular protein molecules.

The lipid bilayer is composed almost entirely of ›hospholipids and cholesterol. One part of the ›hospholipid and the cholesterol molecules is solıble in water, that is, *hydrophilic*, whereas the ›ther part is soluble only in fats, that is, *hydro-›hobic*. The phosphate radical of the phospholipid s hydrophilic, and the fatty acid radicals are fatsoluble. The cholesterol has a hydroxyl radical :hat is water soluble and a steroid nucleus that is fat soluble. Because the hydrophobic portions of both these molecules are repelled by water but are mutually attracted to each other, they have a natural tendency to line up, as illustrated in Figure 2–3, the fatty portions occupying the center of the membrane and the hydrophilic portions projecting to the two surfaces in contact with the surrounding water.

The membrane lipid bilayer is a major barrier impermeable to the usual water-soluble substances such as ions, glucose, urea, and others. On the other hand, fat-soluble substances such as oxygen and alcohols can penetrate this portion of the membrane with ease.

A special feature of the lipid bilayer is that it is a lipid *fluid* and not a solid. Therefore, portions of the membrane can literally flow from one point to another in the membrane. Proteins or other substances dissolved in or floating in the lipid bilayer tend to diffuse to all areas of the cell membrane.

The Cell Membrane Proteins. Figure 2–3 illustrates globular masses floating in the lipid bilayer. These are membrane proteins, most of which are *glycoproteins*. Two types of proteins occur: the *integral proteins* that protrude all the way through the membrane and the *peripheral proteins* that are attached only to the surface of the membrane and do not penetrate.

The integral proteins provide structural *channels* (or *pores*) through which water-soluble substances, especially the ions, can diffuse between the extracellular and intracellular fluid. However, these proteins have selective properties that cause preferential diffusion of some substances more than others. Some of them also act as *carrier proteins* for transporting substances that are too

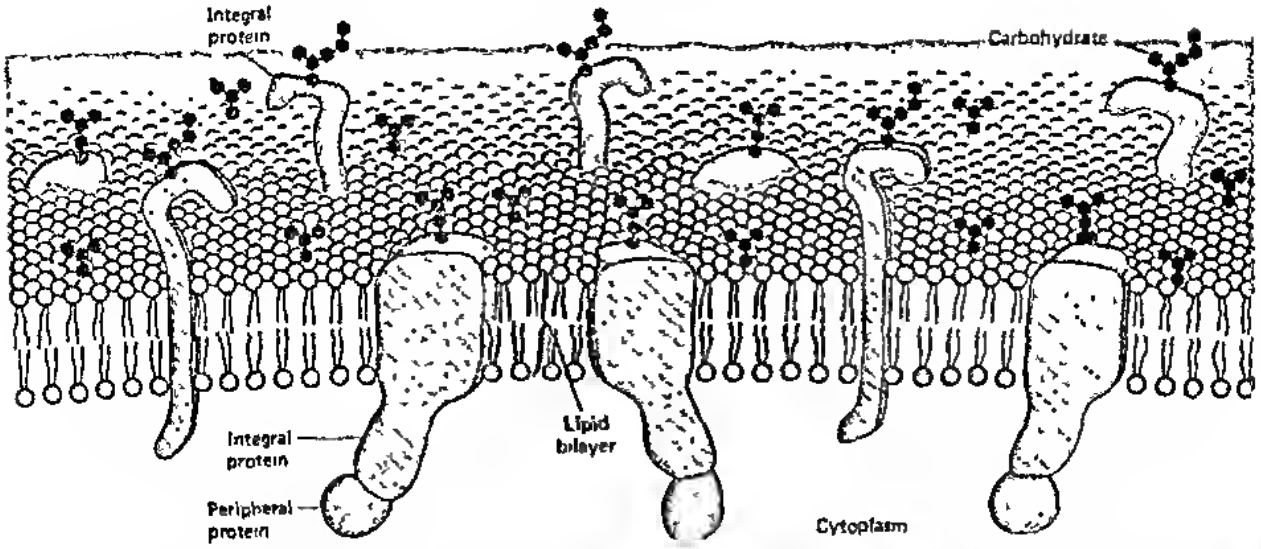


Figure 2–3. Structure of the cell membrane, showing that it is composed mainly from a lipid bilayer but with large numbers of protein molecules protruding through the layer Also, carbohydrate moieties are attached to the protein molecules on the outside of the membrane and additional protein molecules on the inside (from Lodish and Rothman *Sci. Am*, *240*:48, 1979 ©1979 by Scientific American, Inc. All rights reserved)

large to diffuse through the channels; still others act as *enzymes*

The peripheral proteins occur either entirely or almost entirely on the inside of the membrane, and they are normally attached to one of the integral proteins. These peripheral proteins function almost entirely as enzymes

The Membrane Carbohydrates—The Cell "Glycocalyx." The membrane carbohydrotes occur almost invariably in combination with proteins and lipids in the form of *glycoproteins* and *glycolipids*. In fact, most of the integral proteins are glycoproteins, and about one tenth of the lipid molecules are glycolipids The "glyco" portions of these molecules almost invariably protrude to the outside of the cell, dangling outward from the cell surface Many other carbohydrate compounds, called *proteoglycans*, which are mainly carbohydrate substances bound together by small protein cores, are often loosely attached to the outer surface of the cell as well Thus, the entire surface of the cell often has a loose carbohydrate coat called the *glycocalyx*.

These carbohydrate moieties attached to the outer surface of the cell have several important functions· (a) Many of them are negatively charged, which gives most cells an overall negative surface charge that repels other negative objects. (b) The glycocalyx of some cells attaches to the glycocalyx of other cells, thus attaching the cells to each other as well (c) Some carbohydrates act as receptor substances for binding hormones like insulin that stimulate specific types of activity in the cells And (d) some enter into immune reactions, as we shall discuss in Chapter 6.

THE CYTOPLASM AND ITS ORGANELLES

The cytoplasm is filled with both minute a large dispersed particles and organelles rangi in size from a few nanometers to many micro The clear fluid portion of the cytoplasm in whi the particles are dispersed is called *cytosol*; th contains mainly dissolved proteins, electrolyt glucose, and small quantities of phospholipi cholesterol, and esterified fatty acids.

The portion of the cytoplasm immediately b neath the cell membrane frequently contains lar numbers of microfilaments composed of fibril proteins. These provide a semisolid support for t cell membrane. This zone of the cytoplasm is call the *cortex*, or *ectoplasm*. The cytoplasm betwe the cortex and the nuclear membrane is liquef and is called the *endoplasm*.

Among the large dispersed particles in the toplasm are neutral fat globules, glycogen gr ules, ribosomes, secretory granules, and four pecially important organelles—the *endoplasm reticulum*, the *Golgi apparatus*, the *mitochondr* and the *lysosomes*.

The Endoplasmic Reticulum

Figure 2–2 illustrates in the cytoplasm a n work of tubular and flat vesicular structures cal the *endoplasmic reticulum*. The tubules and ve cles all interconnect with each other. Also, th walls are constructed of lipid bilayer membra containing large amounts of proteins, similar the cell membrane. The total surface area of t atructure in some cells—the liver cells, for

MEDICAL COLL

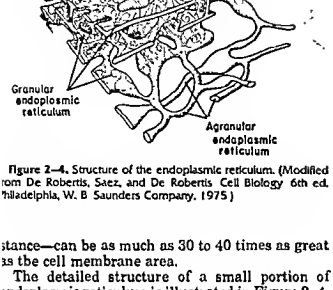


Figure 2–4. Structure of the endoplasmic reticulum. (Modified from De Robertis, Saez, and De Robertis Cell Biology 6th ed. Philadelphia, W. B Saunders Company, 1975)

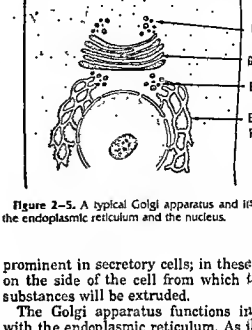


Figure 2–5. A typical Golgi apparatus and its relationship to the endoplasmic reticulum and the nucleus.

stance—can be as much as 30 to 40 times as great as the cell membrane area.

The detailed structure of a small portion of endoplasmic reticulum is illustrated in Figure 2–4. The space inside the tubules and vesicles is filled with *endoplasmic matrix*, a fluid medium that is different from the fluid outside the endoplasmic reticulum. Electron micrographs show that the space insida the endoplasmic reticulum is connected with the space between the two membranes of the double nuclear membrane.

Substances formed in different parts of the cell enter tha space of the endoplasmic reticulum and are then conducted to other parts of the cell. Also, the vast surface area of the reticulum, as well as the many enzyme systems along the surfaces of its membranes, provides the machinery for a major share of the metabolic functions of the cell.

Ribosomes and the Granular Endoplasmic Reticulum. Attached to the outer surfaces of many parts of the endoplasmic reticulum are large numbers of small granular particles called *ribosomes*. Where these are present, the reticulum is frequently called the *granular endoplasmic reticulum*. The ribosomes are composed mainly of ribonucleic acid, which functions in the synthesis of protein in the cells, as discussed later in this and the following chapter.

The Agranular Endoplasmic Reticulum. Part of the endoplasmic reticulum has no attached ribosomes. This part is called the *agranular*, or *smooth, endoplasmic reticulum*. The agranular reticulum functions in the synthesis of lipid substances and in many other enzymatic processes of the cell.

Golgi Apparatus

The Golgi apparatus, illustrated in Figure 2–5, is closely related to the endoplasmic reticulum. It has membranes similar to those of the agranular endoplasmic reticulum. It is usually composed of four or more stacked layers of thin, flat vesicles lying near the nucleus. This apparatus is very prominent in secretory cells; in these it is located on the side of the cell from which the secretory substances will be extruded.

The Golgi apparatus functions in association with the endoplasmic reticulum. As illustrated in Figure 2–5, small "transport vesicles" continually pinch off from the endoplasmic reticulum and shortly thereafter fuse with the Golgi apparatus. In this way substances are transported from the endoplasmic reticulum to the Golgi apparatus. The transported substances are then processed in the Golgi apparatus to form *lysosomes*, *secretory vesicles*, or other cytoplasmic components to be discussed later in the chapter.

The Lysosomes

Another organelle, formed by the Golgi apparatus and dispersed throughout the cytoplasm, is the lysosome. The lysosomes provide an intracellular digestive system that allows the cell to digest and thereby remove unwanted substances and structures, especially damaged or foreign structures, such as bacteria. The lysosome, illustrated in Figure 2–2, is quite different from one cell to another, but it is usually 250 to 750 nanometers in diameter. It is surrounded by a typical lipid bilayer membrane and is filled with large numbers of small granules 5 to 8 nanometers in diameter, which are protein aggregates of hydrolytic (digestive) enzymes. A hydrolytic enzyme is capable of splitting an organic compound into two or more parts by combining hydrogen from a water molecule with part of the compound and by combining the hydroxyl portion of the water molecule with the other part of the compound. For instance, protein is hydrolyzed to form amino acids, and glycogen is hydrolyzed to form glucose. More than 40 different *acid hydrolases* have been found in lysosomes, and the principal substances that they digest are proteins, nucleic acids, mucopolysaccharides, lipids, and glycogen.

Ordinarily, the membrane surrounding the lysosome prevents the enclosed hydrolytic enzymes

from coming in contact with other substances in the cell. However, many different conditions of the cell will break the membranes of some of the lysosomes, allowing release of the enzymes. These enzymes then split the organic substances with which they come in contact into small, highly diffusible substances, such as amino acids and glucose. Some of the more specific functions of lysosomes are discussed later in the chapter.

The Peroxisomes

Peroxisomes are similar physically to lysosomes, but they are different in two important ways. First, they are formed by budding off from the smooth endoplasmic reticulum rather than from the Golgi apparatus. Second, they contain oxidases rather than hydrolases. Several of the oxidases are capable of combining oxygen with hydrogen ions from different intracellular chemicals to form hydrogen peroxide (H_2O_2). The hydrogen peroxide in turn is itself a highly oxidizing substance, and this is used in association with *catalase*, another oxidase enzyme present in large quantities in peroxisomes, to oxidize many substances that might otherwise be poisonous to the cell. For instance, most alcohol that a person drinks is detoxified by the peroxisomes of the liver cells in this manner. And the hydrogen peroxide-catalase-oxidizing mechanism is also used for specific functional purposes in the cell, such as catalyzing the breakdown of fatty acids into acetyl-CoA that is then to be used for energy by the cell.

Secretory Vesicles

One of the important functions of many cells is secretion of special substances. Almost all such secretory substances are formed by the endoplasmic reticulum-Golgi apparatus system and are then released from the Golgi apparatus into the cytoplasm inside storage vesicles called *secretory vesicles* or *secretory granules*. Figure 2-6 illustrates typical secretory vesicles inside pancreatic acinar cells that store protein enzymes; the enzymes will be secreted later through the outer cell membrane into the pancreatic duct.

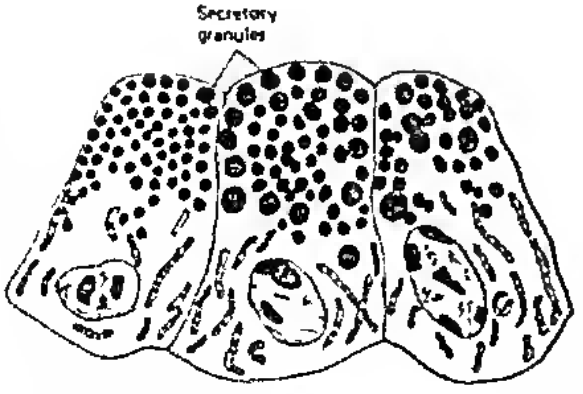


Figure 2-6. Secretory granules in acinar cells of the pancreas.

Other Cytoplasmic Structures and Organelles

Aside from the membranous organelles of the cytoplasm—the endoplasmic reticulum, the Golgi apparatus, the lysosomes, and the peroxisomes—the cytoplasm also contains several other structures or organelles. For instance, usually equally as many *ribosomes* float freely in the cytoplasm as attach to the granular endoplasmic reticulum. These function for the formation of cytoplasmic proteins, in contradistinction to the endoplasmic reticulum proteins that are formed by the ribosomes attached to the endoplasmic reticulum and pass to the interior of the endoplasmic chamber. Most of the other cytoplasmic structures are constructed of microfilaments, or microtubules, as follows:

Microfilaments and Microtubular Structures in the Cell. The fibrillar proteins of the cell cytoplasm are usually organized into microfilaments or microtubules. These originate as precursor protein molecules synthesized by the ribosomes that float freely in the cytoplasm. At first they are present in dissolved form in the cytoplasm. Then they polymerize to form *microfilaments*. We have already pointed out that large numbers of microfilaments frequently occur in the outer zone of the cytoplasm, the zone called the *ectoplasm*, to form an elastic support for the cell membrane. Also, in muscle cells microfilaments are organized into a special contractile machine that is the basis of muscle contraction throughout the body, as will be discussed in detail in Chapter 11.

Microfilaments are also frequently organized into tubular structures, the *microtubules*. Almost invariably these contain 13 microfilaments lying parallel and in a circle to form a long hollow cylinder about 25 nanometers in diameter and up to many microns in length. These are often arranged in bundles that gives them, *en masse*, considerable structural strength. However, microtubules are stiff structures that break if bent too severely. Figure 2-7 illustrates typical microtubules that were teased from the flagellum of a sperm.

Another example of microtubules is the tubular mechanical structure of cilia that gives them structural strength, radiating upward from the cell cytoplasm to the tip of the cilium. Also, the *centrioles* and the *mitotic spindle* of the mitosing cell are both composed of stiff microtubules.

Thus, a primary function of microtubules is to act as a *cytoskeleton*, providing rigid physical structures for certain parts of cells. Also, it has been noted that the cytoplasm often *streams* in the vicinity of microtubules, which might result from movement of arms that project outward from the microtubules.

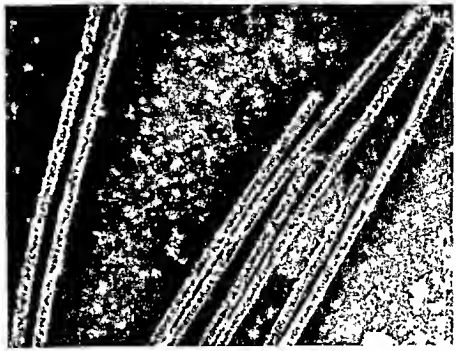


Figure 2–7. Microtubules teased from the igellum of a sperm. (From Poner, Ciba Foun- ation Symposium. Principles of Biomolecular rganization. Boston, Little, Brown and Com- , 1966.)

he Mitochondria

The mitochondria are called the "powerhouses" f the cell. Without them the cells would be unable o extract significant amounts of energy from the utrients and oxygen, and as a consequence essen- ially all cellular functions would cease. As illus- rated in Figure 2–2, these organelles are present n essentially all portions of the cytoplasm, but he number per cell varies from less than a undred to several thousand, depending upon the mount of energy required by each cell. Further- nore, the mitochondria are concentrated in those ortions of the cell that are responsible for the najor share of its energy metabolism. Mitochon- dria are also very variable in size and shape; some are only a few hundred millimicrons in diameter and globular in shape while others are as large as 1 micron in diameter, as long as 7 microns, and filamentous in shape.

The basic structure of the mitochondrion is il- lustrated in Figure 2–8, which shows it to be composed mainly of two lipid bilayer-protein mem- branes: an *outer membrane* and an *inner mem- brane*. Many infoldings of the inner membrane form *shelves* onto which oxidative enzymes are attached. In addition, the inner cavity of the mi- tochondrion is filled with a gel *matrix* containing large quantities of dissolved enzymes that are necessary for extracting energy from nutrients. These enzymes operate in association with the oxidative enzymes on the shelves to cause oxida- tion of the nutrients, thereby forming carbon diox- ide and water. The liberated energy is used to synthesize a high-energy substance called *adeno- sine triphosphate (ATP)*. ATP is then transported out of the mitochondrion, and it diffuses through- out the cell to release its energy wherever it is needed for performing cellular functions. The de- tails of ATP formation by the mitochondrion are given in Chapter 67, and some of the important functions of ATP in the cell are introduced later in this chapter.

Mitochondria are self-replicative, which means that one mitochondrion can form a second one, a third one, and so on, whenever there is need in the cell for increased amounts of ATP. Indeed, the mitochondria contain a special type of deoxyribo- nucleic acid similar to that found in the nucleus. In the following chapter we shall see that deoxy- ribonucleic acid is the basic substance that controls replication of all the constituents of the cell. It has now been demonstrated that this substance plays a similar role in the mitochondrion, but not en- tirely so because in the mitochondrion replication process, many proteins and lipids that have al-

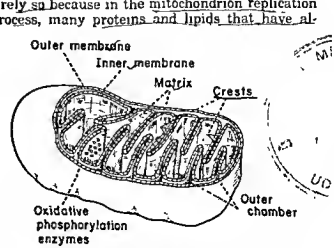


Figure 2–8. Structure of a mitochondrion. (Modified from De Robertis, Saez, and De Robertis. Cell Biology. 6th ed. Philadelphia: W. B. Saunders Company, 1975.)

ready been formed in the cytoplasm are incorporated into the mitochondria as they enlarge and then bud off to form new mitochondria.

THE NUCLEUS

The nucleus is the control center of the cell. It controls both the chemical reactions that occur in the cell and reproduction of the cell. Briefly, the nucleus contains large quantities of *deoxyribonucleic acid*, which we have called *genes* for many years The genes determine the characteristics of the protein enzymes of the cytoplasm, and in this way control cytoplasmic activities. To control reproduction, the genes first reproduce themselves, and after this is accomplished the cell splits by a special process called *mitosis* to form two daughter cells, each of which receives one of the two sets of genes. All these activities of the nucleus are considered in detail in the following chapter.

The appearance of the nucleus under the microscope does not give much of a clue to the mechanisms by which it performs its control activities. Figure 2–9 illustrates the light microscopic appearance of the interphase nucleus (period between mitoses), showing darkly staining *chromatin material* throughout the nucleoplasm. During mitosis, the chromatin material becomes readily identifiable as part of the highly atructured *chromasomes*, which can be seen easily with the light microscope. Even during the interphase of cellular activity the chromatin material is still organized into fibrillar chromosomal structures, but this is difficult to see except in a few types of cells.

The Nuclear Envelope

The nuclear envelope is frequently called the *nuclear membrane*. However, it is actually two separate membranes, one inside the other. The outer membrane is continuous with the endoplasmic reticulum, and the space between the two nuclear membranes is also continuous with the compartment inside the endoplasmic reticulum.

The nuclear envelope is penetrated by several thausand *nuclear pores*. These are very large, most 100 nanometers in diameter. However, la complexes of protein molecules are attac around the edges of the pores so that the cent area of the pore is only about 9 nanometers diameter. Even this size is large enough to all protein molecules as large as 44,000 molecu weight to pass through with relative ease proteins with molecular weight less than 15, to pass extremely rapidly.

Nucleoli

The nuclei of many cells contain one or m lightly staining structures called nucleoli. nucleolus, unlike most of the organelles that have discussed, does not have a limiting m brane. Instead, it is simply a structure that tains a large amount of *ribonucleic acid*, and teins of the types found in ribosomes. nucleolus becomes considerably enlarged whe cell is actively synthesizing proteins. The gene five separate chromosome pairs synthesize the bonucleic acid and then store it in the nucleo beginning with a loose fibrillar RNA that l condenses to form granular "subunits" of r somes These in turn are transported through nuclear membrane pores into the cytoplasm wh they assemble together to form the "mature" bosomes that play an essential role for the for tion of proteins, either in the cytosol or in ass ation with the endoplasmic reticulum, as we s discuss more fully in the following chapter.

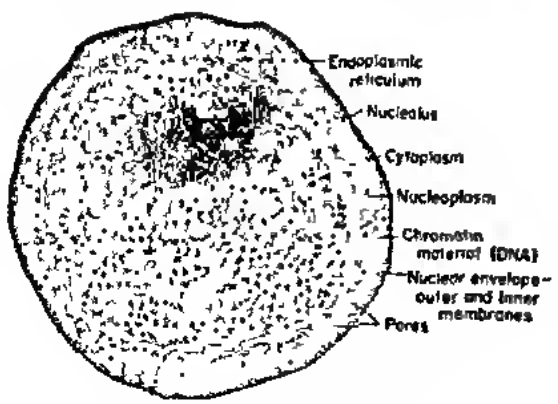


Figure 2–9. Structure of the nucleus.

COMPARISON OF THE ANIMAL CELL WITH PRECELLULAR FORMS OF LIFE

Many of us think of the cell as the lowest level of However, the cell is a very complicated organism, w probably required many hundreds of million years develop after the earliest form of life, an organ similar to the present-day *virus*, first appeared on ea Figure 2–10 illustrates the relative sizes of the small

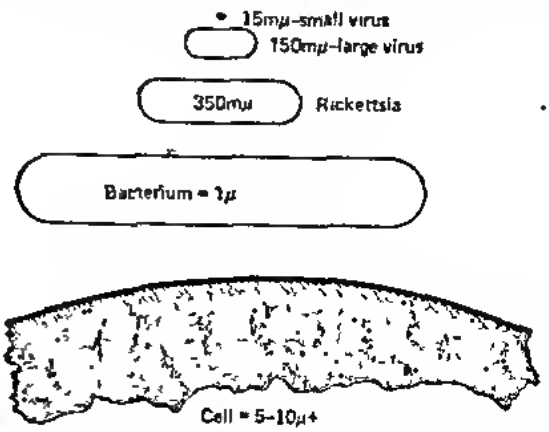


Figure 2–10. Comparison of sizes of precellular organisms that of the average cell in the human body

known virus, a large virus, a *rickettsia*, a *bacterium*, and a nucleated cell, showing that the cell has a diameter about 1000 times that of the smallest virus, and, *therefore, a volume about 1 billion times that of the* smallest virus. Correspondingly, the functions and an- organization of the cell are also far more complex than those of the virus.

The essential life-giving constituent of the very small virus is a *nucleic acid* embedded in a coat of protein. This nucleic acid is similar to that of the cell, and it is capable of reproducing itself if appropriate surrounding conditions are available. Thus, the virus is capable of propagating its lineage from generation to generation, and, therefore, is a living structure in the same way that the cell and the human being are living structures.

As life evolved, other chemicals besides nucleic acid and simple proteins became integral parts of the organism, and specialized functions began to develop in different parts of the virus. A membrane formed around the virus, and inside the membrane a fluid matrix appeared. Specialized chemicals developed inside the matrix to perform special functions; many protein enzymes appeared which were capable of catalyzing chem- reactions, and, therefore, of controlling the organism's activities.

In still later stages, particularly in the rickettsial and bacterial stages, *organelles* developed inside the organism, representing physical structures of chemical aggregates that perform functions in a more efficient manner than can be achieved by dispersed chemicals throughout the fluid matrix. And, finally, in the nucleated cell, still more complex organelles developed, the most important of which is the *nucleus* itself. The nucleus distinguishes *this type of cell from all lower forms of life, this structure* provides a control center for all cellular activities, and it also provides for very exact reproduction of new cells generation after generation, each new cell having essentially the same structure as its progenitor.

FUNCTIONAL SYSTEMS OF THE CELL

In the remainder of this chapter we will discuss several representative functional systems of the cell that make it a living organism.

INGESTION BY THE CELL—ENDOCYTOSIS

If a cell is to live and grow, it must obtain nutrients and other substances from the surrounding fluids. Substances can pass through a cell membrane in three separate ways: (1) by *diffusion* through the pores in the membrane or through the membrane matrix itself; (2) by *active transport* through the membrane, a mechanism in which enzyme systems and special carrier proteins transport the substances through the membrane; and (3) by *endocytosis*, a mechanism by which the membrane actually engulfs particulate matter or extracellular fluid and its contents. The important subject of transport of substances by diffusion and active transport, the means by which most nutrients and other substances enter the cell, will be considered in detail in Chapter 9. Endocytosis is a specialized cellular function that merits mention here.

Endocytosis begins by the formation of small vesicles at the cell membrane; these vesicles then pinch off to the interior of the cell and float freely in the cytoplasm. The two principal forms of endocytosis are called *pinocytosis* and *phagocytosis*. Pinocytosis means the formation of extremely small vesicles containing extracellular fluid. Phagocytosis means ingestion of large particles, such as bacteria, cells, or portions of degenerating tissue.

Pinocytosis. Pinocytosis occurs continually at the cell membranes of most cells, but especially rapidly in some cells. For instance, it occurs so rapidly in macrophages that about 3 per cent of the total macrophage membrane is engulfed in the form of vesicles each minute. Even so, the pinocytic vesicles are so small, usually only 100 to 200 nanometers in diameter, that most of their characteristics can be discerned only with the electron microscope.

Pinocytosis is the only means by which some very large macromolecules, such as most protein molecules, can enter cells. In fact, the rate at which pinocytic vesicles form usually is enhanced when such macromolecules attach to the cell membrane.

Figure 2–11 illustrates the successive steps of pinocytosis, showing three molecules of protein attaching to the membrane. These molecules usually attach *to receptors in the membrane that* are specific for the types of proteins that are to be absorbed, and these receptors generally are concentrated in small pits in the cell membrane, called *coated pits*. The inside surfaces of these pits, in turn, are coated with a dense material containing contractile filaments. Once the protein molecules have bound with the receptors, the surface properties of the membrane change in such a way that the entire pit invaginates inward and its borders close over the attached proteins, as well as over a small amount of extracellular fluid. Immediately

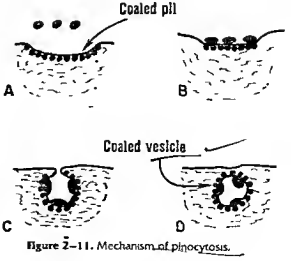


Figure 2–11. Mechanism of pinocytosis.

thereafter, the invaginated portion of the membrane breaks away from the surface of the cell, forming a *pinocytic vesicle*.

What causes the cell membrane to go through the necessary contortions for forming pinocytic vesicles remains mainly a mystery. However, this process requires energy from within the cell; this is supplied by adenosine triphosphate, a high energy substance that will be discussed later in the chapter. Also, it requires the presence of calcium ions in the extracellular fluid, which probably react with the contractile microfilaments beneath the coated pits to provide the force for pinching the vesicles away from the cell membrane.

Phagocytosis. Phagocytosis occurs in much the same way as pinocytosis, except that it involves large particulate matter rather than macromolecules. Only certain cells have the capability of phagocytosis, most notably the tissue macrophages and some of the white blood cells. A postulated cause of phagocytosis is the presence of positive electrical charges on the surfaces of the phagocytized particles. Most normal particles in the extracellular fluid are negatively charged, the same as the outsides of the macrophages and white blood cells. Therefore, these particles are repelled from the cell surfaces. On the other hand, damaged tissues and bacteria that have been attacked by antibodies (a process called *opsonization*, which will be discussed in Chapters 5 and 6) will often acquire positive charges or other surface properties that will allow them to be phagocytized.

Phagocytosis occurs in the following steps: First, the cell membrane attaches to the surface of the object. Second, the edges of the membrane around the point of attachment suddenly evaginate outward and completely engulf the object within a fraction of a second. Third, the membrane contracts and pulls the object to the interior. And, fourth, the phagocytic vesicle pinches off to the interior of the cell in the same way that pinocytic vesicles are formed.

DIGESTION OF FOREIGN SUBSTANCES IN THE CELL—FUNCTION OF THE LYSOSOMES

Almost immediately after a pinocytic or phagocytic vesicle appears inside a cell, one or more lysosomes become attached to the vesicle and [illegible] their hydrolases into the vesicle, as illustrated in Figure 2–12. Thus, a *digestive vesicle* is [illegible] in which the hydrolases begin hydrolyzing [illegible] proteins, glycogen, nucleic acids, mucopolysaccharides, and other substances in the vesicle. The products of digestion are small molecules of amino acids, glucose, phosphates, and so forth that can then diffuse through the membrane of the vesicle into the cytoplasm. What is left of the digestive vesicle, called the *residual body*, represents the undigestible substances. In most instances this is finally excreted through the cell membrane by a

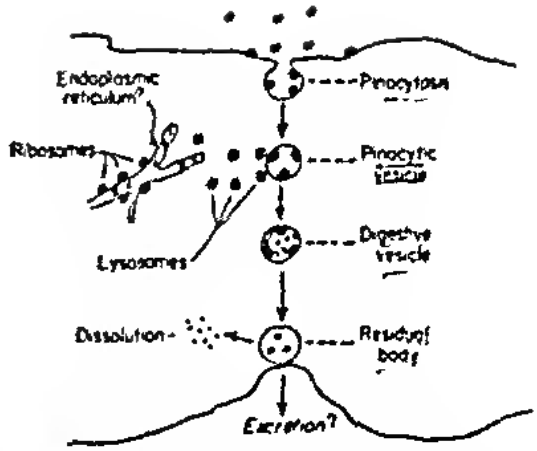


Figure 2–12. Digestion of substances in pinocytic vesicles by enzymes derived from lysosomes.

process called *exocytosis*, which is essentially the opposite of endocytosis.

Thus, the lysosomes may be called the *digestive organs* of the cells.

Regression of Tissues and Autolysis of Cells. Often, tissues of the body regress to smaller size. For instance, this occurs in the uterus following pregnancy, in muscles during long periods of inactivity, and in mammary glands at the end of the period of lactation. Lysosomes are responsible for much of this regression. However, the mechanism by which lack of activity in a tissue causes the lysosomes to increase their activity is still unknown.

Another very special role of the lysosomes is the removal of damaged cells or damaged portions of cells from tissues—cells damaged by heat, cold, trauma, chemicals, or any other factor. Damage to the cell causes lysosomes to rupture, and the released hydrolases begin immediately to digest the surrounding organic substances. If the damage is slight, only a portion of the cell will be removed, followed by repair of the cell. However, if the damage is severe the entire cell will be digested, a process called *autolysis*. In this way, the cell is completely removed and a new cell of the same type ordinarily is formed by mitotic reproduction of an adjacent cell to take the place of the old one.

The lysosomes also contain bactericidal agents that can kill phagocytized bacteria before they can cause cellular damage. These agents include *lysozyme* that dissolves the bacterial cell membrane, *lysoferrin* that binds iron and other metals that are essential for bacterial growth, and acid at a pH of about 5.0 that activates the hydrolases and also inactivates some of the bacterial metabolic systems.

Also in lysosomes are stored enzymes that can dissolute lipid droplets and glycogen granules, making the lipid and glycogen available for use elsewhere in the cell or elsewhere in the body. In the absence of these enzymes, which results from occasional genetic disorders, extreme quantities of

pids or of glycogen often accumulate in the cells f many organs, especially the liver, and lead to arly death.

YNTHESIS AND FORMATION OF CELLULAR TRUCTURES BY THE ENDOPLASMIC ETICULUM AND THE GOLGI APPARATUS

The extensiveness of the endoplasmic reticulum nd the Golgi apparatus, especially in secretory ells, has already been emphasized. These two tructures are made primarily of lipid bilayer embranes, and their walls are literally loaded ith protein enzymes that catalyze the synthesis f many of the substances required by the cell.

In general, most of the synthesis begins in the ndoplasmic reticulum, but the products formed in he endoplasmic reticulum are then passed on to he Golgi apparatus where they are further processed prior to release into the cell. But, first, let us note the specific products that are synthesized n the specific portions of the endoplasmic reticulum and the Golgi apparatus.

Formation of Proteins by the Granular Endoplasmic Reticulum. The granular endoplasmic eticulum is characterized by the presence of large umbers of ribosomes attached to the outer surfaces of the reticulum membrane. As we shall liscuss in the following chapter, protein molecules are synthesized within the structures of the ribosomes. Furthermore, the ribosomes extrude many of the synthesized protein molecules not into the ytosol but instead through the endoplasmic reticulum wall into the endoplasmic matrix.

Almost as rapidly as the protein molecules enter he endoplasmic matrix, enzymes in the endoplasmic reticulum wall cause rapid changes in hese molecules. First, almost all of them are mmediately glycosylated, that is, conjugated with carbohydrate moieties, to form *glycoproteins*. Therefore, essentially all the endoplasmic proteins are glycoproteins, in contrast to the proteins that are formed by the ribosomes in the cytosol, which are mainly free proteins. Second, the proteins are cross-linked and folded to form more compact molecules.

Synthesis of Lipids by the Endoplasmic Reticulum, Especially by the Smooth Endoplasmic Reticulum. The endoplasmic reticulum also synthesizes lipids, especially phospholipids and cholesterol. These are rapidly incorporated into the lipid bilayer of the endoplasmic reticulum itself, thus causing the endoplasmic reticulum to grow continually. This occurs mainly in the smooth portion of the endoplasmic reticulum.

To keep the endoplasmic reticulum from growing beyond the limits of the cell, however, small vesicles called *endoplasmic reticulum vesicles*, or *transport vesicles*, continually break away from the smooth reticulum; we shall see later that most of these vesicles migrate rapidly to the Golgi apparatus.

Other Functions of the Endoplasmic Reticulum. Other significant functions of the endoplasmic reticulum, again especially the smooth reticulum, are:

(1) It contains the enzymes that control glycogen breakdown when glycogen is to be used for energy.

(2) It contains a vast number of enzymes that are capable of detoxifying substances that are damaging to the cell, such as drugs. It achieves this by coagulation, oxidation, hydrolysis, conjugation with glycuronic acid, and in other ways.

Synthetic Functions of the Golgi Apparatus. Though the major function of the Golgi apparatus is to process substances already formed in the endoplasmic reticulum, it also has the capability of synthesizing certain carbohydrates that cannot be formed in the endoplasmic reticulum. This is especially true of sialic acid and galactose. In addition, it can cause the formation of very large saccharide polymers bound with only small amounts of protein; the most important of these are hyaluronic acid and chondroitin sulfate. A few of the many functions of hyaluronic acid and chondroitin sulfate in the body are: (1) They are the major components of proteoglycans secreted in mucus and other glandular secretions. (2) They are the major components of the ground substance in the interstitial spaces, acting as a filler between collagen fibers and cells. (3) They are principal components of the organic matrix in both cartilage and bone.

Processing of Endoplasmic Secretions by the Golgi Apparatus—Formation of Intracellular Vesicles. Figure 2–13 summarizes the major functions of the endoplasmic reticulum and Golgi apparatus. As substances are formed in the endoplasmic reticulum, especially the proteins, they are transported through the tubules toward the portions of the smooth endoplasmic reticulum that lie nearest the Golgi apparatus. At this point small "transport" vesicles of smooth endoplasmic reticulum continually break away and diffuse to the *proximal layers* of the Golgi apparatus. Inside these vesicles are the synthesized proteins and

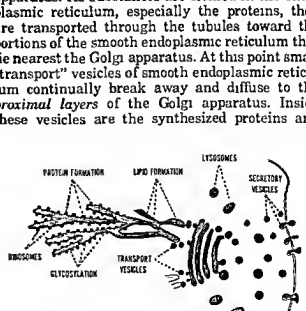


Figure 2–13. Formation of proteins, lipids, and cellular vesicles by the endoplasmic reticulum and Golgi apparatus.

other products. These vesicles instantly fuse with the Golgi apparatus and empty their contained substances into the vesicular spaces of the Golgi apparatus Here, additional carbohydrate moieties are added to the secretions Also, a most important function of the Golgi apparatus is to compact the endoplasmic reticular secretions into highly concentrated packets. As the secretions pass toward the distal layers of the Golgi apparatus, the compaction and processing proceed; finally, at the distal layer both small and large vesicles continually break away from the Golgi apparatus, carrying with them the compacted secretory substances, and they then diffuse throughout the rell.

To give one an idea of the timing of these processes: when a glandular cell is bathed in radioactive amino acids, newly formed radioactive protein molecules can be detected in the granular endoplasmic reticulum within 3 to 5 minutes Within 20 minutes the newly formed proteins are present in the Golgi apparatus, and within 1 to 2 hours radioactive proteins are secreted from the surface of the cell.

Types of Vesicles Formed by the Golgi Apparatus—Secretory Vesicles and Lysosomes. In a highly secretory cell, the vesicles that are farmed by the Golgi apparotus are mainly *secretory vesicles*, containing especially the protein substances that are to be secreted through the surfare of the cell These vesicles diffuse to the cell membrane, then fuse with it and empty their substances to the exterior by a mechanism called *exocytosis*, which is essentially the opposite of endocytosis. Exocytosis, in most cases, is stimulated by entry of calcium ions into the cell, the calcium ions interact with the vesicular membrane, in some way not understood, to cause its fusion with the cell membrane.

On the other hand, some of the vesicles are destined for intracellular use. For instanre, specialized portions of the Golgi apparatus form the *lysosomes* that have already been discussed It is believed that the membranes of these specialized portions contain chemical receptors that cause acid hydrolases to attach. In this way these enzymes are concentrated and then released from the Golgi apparatus in the form of the lysosomal vesicles.

Another type of vesicle formed in the same way as the lysosomes are the *peroxisomes* However, it is believed that these are formed by the smooth endoplasmic reticulum, along with formation of the transport vesicles, rather than by the Golgi apparatus Here again, special receptors in the endoplasmic reticulum membrane probably attract and attach the oxidative enzymes that are to be released in concentrated form in the peroxisomes.

Use of Intracellular Vesicles to Replenish Cellular Membranes. Many of the vesicles finally fuse with the cell membrane or with the membranes of other intracellular stcuctures such as the mitochondria and even the endoplasmic reticulum itself. This obviously increases [illegible] of these membranes and thereby repleni[illegible] membranes as they themselves are destroyed [illegible] instance, the cell membrane loses much of [illegible] substance every time it forms a phagocytic [illegible] pinocytic vesicle, and it is vesicles from the Golgi apparatus that continually replenish the cell membrane

Thus, in summary, the membranous system [illegible] the endoplasmic reticulum and the Golgi [illegible] tus represents a highly metabolic organ capable [illegible] forming both new cellular structures and [illegible] substances to be extruded from the cell.

EXTRACTION OF ENERGY FROM NUTRIENTS—FUNCTION OF THE MITOCHONDRIA

The principal substances from which cells [illegible] tract energy are oxygen and one or more of [illegible] foodstuffs—carbohydrates, fats, and proteins [illegible] the human body essentially all carbohydrates [illegible] converted into *glucose* before they reach the [illegible] the proteins are converted into *amino acids*, [illegible] the fats are converted into *fatty acids*. Figure 2-[illegible] shows oxygen and the foodstuffs—glucose, fa[illegible] acids, and amino acids—all entering the cell. [illegible] side the cell, the foodstuffs react chemically w[illegible] the oxygen under the influence of various enzy[illegible] that control their rates of reactions and chan[illegible] the energy that is released in the proper directi[illegible]

Almost all these oxidative reactions occur ins[illegible] the mitochondria, and the energy that is relea[illegible] is used to form adenosine triphosphate. Then, [illegible] adenosine triphosphate, not the original foodst[illegible] themselves, is used throughout the cell to energ[illegible] almost all the intracellular metabolic reactio[illegible] Thus, the adenosine triphosphate is an intra[illegible] lular storehouse of energy.

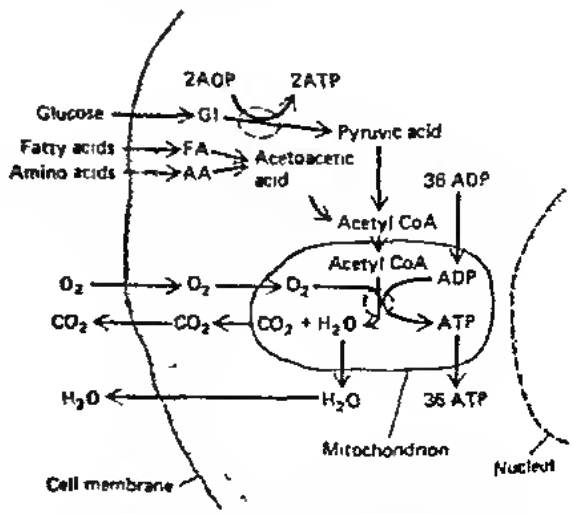
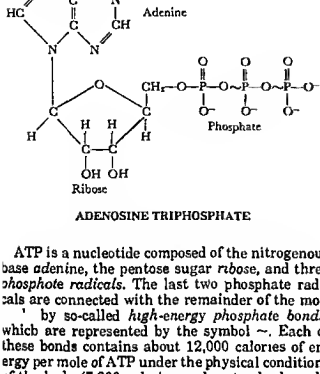


Figure 2–14. Formation of adenosine triphosphate in the [illegible] showing that most of the ATP is formed in the mitochondria.

Functional Characteristics of Adenosine Triphosphate (ATP). The formula for adenosine triphosphate, generally called simply ATP, is:



ADENOSINE TRIPHOSPHATE

ATP is a nucleotide composed of the nitrogenous base *adenine*, the pentose sugar *ribose*, and three *phosphate radicals*. The last two phosphate radicals are connected with the remainder of the molecule by so-called *high-energy phosphate bonds*, which are represented by the symbol ~. Each of these bonds contains about 12,000 calories of energy per mole of ATP under the physical conditions of the body (7,300 calories under standard conditions), which is much greater than the energy stored in the average chemical bond to other organic compounds, thus giving rise to the term "high-energy" bond. Furthermore, the high-energy phosphate bond is very labile so that it can be split instantly on demand whenever energy is required to promote other cellular reactions.

When ATP releases its energy, a phosphoric acid radical is split away, and *adenosine diphosphate (ADP)* is formed. Then, energy derived from the cellular nutrients causes the ADP and phosphoric acid to recombine to form new ATP, the entire process continuing over and over again. For these reasons, ATP has been called the *energy currency* of the cell, for it can be spent and remade again and again.

Chemical Processes in the Formation of ATP—Role of the Mitochondria. On entry into the cells, glucose is subjected to enzymes in the cytoplasm that convert it into *pyruvic acid* (a process called *glycolysis*). A small amount of ADP is changed into ATP by energy released during this conversion, but this amount accounts for less than 5 per cent of the overall energy metabolism of the cell.

By far the major portion of the ATP formed in the cell is formed in the mitochondria. The pyruvic and fatty acids and most of the amino acids are all converted into the compound *acetyl·CoA* in the matrix of the mitochondrion. This substance, in turn, is acted upon by a series of enzymes and undergoes dissolution in a sequence of chemical reactions called the *citric acid cycle*, or *Krebs cycle*. These chemical reactions will be explained in detail in Chapter 67.

In the citric acid cycle, acetyl·CoA is split into its component parts, hydrogen atoms and carbon dioxide. The carbon dioxide, in turn, diffuses out of the mitochondria and eventually out of the cell. The hydrogen atoms combine with carrier substances and are carried to the surfaces of the shelves that protrude into the mitochondria, shown in Figure 2–8. Attached to these shelves are the so-called *oxidative enzymes* and also protruding globules of *ATP synthetase*, the enzyme that catalyzes the conversion of ADP to ATP. The oxidative enzymes, by a series of sequential reactions, cause the hydrogen atoms to combine with oxygen. The enzymes are arranged on the surfaces of the shelves in such a way that the products of one chemical reaction are immediately relayed to the next enzyme, then to the next, and so on until the complete sequence of reactions has taken place. During the course of these reactions, the energy released from the combination of hydrogen with oxygen is used to activate the ATP synthetase and drive the reaction to manufacture tremendous quantities of ATP from ADP. The ATP is then transported out of the mitochondrion into all parts of the cytoplasm and nucleoplasm where its energy is used to energize the functions of the cell.

The details of ATP formation end many of its metabolic functions in the body will be presented in Chapters 67 through 71.

Uses of ATP for Cellular Function. ATP is used to promote three major categories of cellular functions: (1) *membrane transport*, (2) *synthesis of chemical compounds* throughout the cell, and (3) *mechanical work*. These three different uses of ATP are illustrated in Figure 2–15: (a) to supply

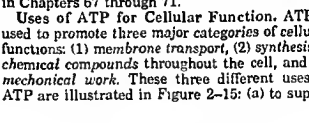


Figure 2–15. Use of adenosine triphosphate to provide energy for three major cellular functions. (1) membrane transport, (2) protein synthesis, and (3) muscle contraction.

energy for the transport of sodium through the cell membrane, (b) to promote protein synthesis by the ribosomes, and (c) to supply the energy needed during muscle contraction

In addition to membrane transport of sodium, energy from ATP is required for transport of potassium ions and, in certain cells, calcium ions, phosphate ions, chloride ions, urate ions, hydrogen ions, and still many other special substances. Membrane transport is so important to cellular function that some cells, the renal tubular cells for instance, utilize as much as 80 per cent of the ATP formed in the cells for this purpose alone.

In addition to synthesizing proteins, cells also synthesize phospholipids, cholesterol, purines, pyrimidines, and a great host of other substances. Synthesis of almost any chemical compound requires energy. For instance, a single protein molecule might be composed of as many as several thousand amino acids attached to each other by peptide linkages; the formation of each of these linkages requires the breakdown of three high energy bonds; thus many thousand ATP molecules must release their energy as each protein molecule is formed Indeed, some cells utilize as much as 75 per cent of all the ATP formed in the cell simply to synthesize now chemical compounds; this is particularly true during the growth phase of cells.

The final major use of ATP is to supply energy for special cells to perform mechanical work. We shall see in Chapter 11 that each contraction of a muscle fiber requires expenditure of tremendous quantities of ATP Other cells perform mechanical work in two additional ways, by *ciliary* or *ameboid motion*, both of which will be described later in this chapter. The source of energy for all these types of mechanical work is ATP

In summary, therefore, ATP is always available to release its energy rapidly and almost explosively wherever in the cell it is needed. To replace the ATP used by the cell, other much slower chemical reactions break down carbohydrates, fats, and proteins and use the energy derived from these to form new ATP. About 95 per cent of this ATP is formed in the mitochondria, which accounts for the name given to the mitochondria, the "powerhouses" of the cell.

LOCOMOTION OF CELLS

By far the most important type of cell movement that rurs in the body is that of the specialized muscle cells in skeletal, cardiac, and smooth muscle, which comprise almost 50 per cent of the entire body mass The specialized functions of these cells will be discussed in Chapters 11 through 13 However, two other types of movement occur in other cells, *ameboid locomotion* and *ciliary movement*.

Ameboid Locomotion. Ameboid locomotion means movement of an entire cell in relation to its surroundings, such as the movement of white blood cells through tissues However, it has received its name from the fact that amebae move in this manner and have provid the best tool for studying the phenomenon. Typical ameboid locomotion begins with protrusion of a *pe dopodium* from one end of the cell. The pseudopodiu projects far out, away from the cell body, and then t remainder of the cell moves toward the pseudopodiu Figure 2–16 illustrates this process, showing an el gated cell the right-hand end of which is a protrudi pseudopodium The membrane of this end of the cell continually moving forward, and the membrane at t left-hand end of the cell is continually following alo as the cell moves

It is believed that ameboid locomotion is caused the following way The outer portion of the cytoplasm in a *gel* state and is called the *ectoplasm*, whereas t central portion of the cytoplasm is in a *sol* state end called *endoplasm* In the gel are numerous microfi ments composed of *actin*, and also present are ma *myosin molecules* that can interact with the actin fi ments to cause contraction, the same as occurs in musc These filaments contract in the presence of ATP a calcium ions, and normally there is a continual tende for the ectoplasm to contract However, in response an external chemical or physical stimulus the ectopla at one end of the cell becomes thin, causing a pseu podium to bulge outward in the direction of the "che tactic" source Thus, the pseudopodium moves forwa At the same time, the ectoplasm at the opposite end the cell contracts, which pushes the endoplasm into t pseudopodium and causes it to extend even furth Also, the membrane of the pseudopodium often becom attached to some of the surrounding tissue elements that when the ectoplasm in the remainder of the c contracts, this literally pulls the main body of the c toward the pseudopodium

Thus, it is very easy to understand how a sin pseudopodium can move forward, then pull the rema der of the cell toward it But once this has occurred, has been difficult to understand how the cell can c tinue to move indefinitely in the direction of an exte ing pseudopodium, because soon all of the cell would in the pseudopodium itself The explanation for t appears to be the following As the ectoplasm in t main body of the cell contracts and the size of the c body becomes smaller, the ectoplasmic layer at t becomes much thicker Then, the inner surface of t thickened ectoplasm loses its gel-like character, a pr ess called *solation*, and provides a continuing amount new endoplasm in the body of the cell. Next, the n

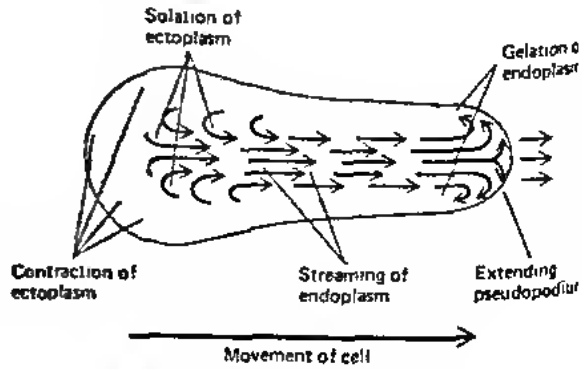


Figure 2–16. Ameboid motion by a cell

doplasm "streams" forward into the pseudopodium, tending it still farther. And finally, the endoplasm jacent to the side walls of the pseudopodium now anges from the fluid state into the gel state, a process lled *gelation*. This thickens the ectoplasm along the le walls of the pseudopodium. By this time the main dy of the cell, including its nucleus and other internal ganelles, is also entering the base of what had been e old pseudopodium. Therefore, what had been the le walls of the pseudopodium now become the walls of e cell body, and the entire process begins once ain—contraction and solation of the ectoplasm in the y, streaming of endoplasm toward the pseudopodium, lation of the side walls of the pseudopodium to form ore ectoplasm, and progressive movement of the cell ward the tip of the pseudopodium

Types of Cells That Exhibit Ameboid Locomotion. ie most common cells exhibiting ameboid locomotion the human body are the *white blood cells* moving out the blood into the tissues in the form of *tissue acrophages* or *microphages*. However, many other pes of cells can move by ameboid locomotion under rtain circumstances. For instance, fibroblasts will ove into any damaged area to help repair the damage, id even some of the germinal cells of the skin, though dinarily completely sessile cells, will move toward a it area to repair the rent. Finally, cell locomotion is pecially important in the development of the fetus, for nbryonic cells often migrate long distances from the nmordial sites of origin to new areas during the delopment of special structures.

Control of Ameboid Locomotion— "Chemotaxis." he most important factor that usually initiates ameoid locomotion, presumably by causing electrical potenal changes in the cell membrane or by the entry of ns into the cell—calcium ions as a possibility—is the ppearance of certain chemical substances in the tissues. his phenomenon is called *chemotaxis*, and the chemical ubstance that causes it to occur is called a *chemotactic ubstance*. Most cells that exhibit ameboid locomotion iove toward the source of the chemotactic subtance—that is, from an area of lower concentration ward an area of higher concentration—which is called *ostive chemotaxis*. However, some cells move away rom the source, which is called *negative chemotaxis*.

ILIA AND CILIARY MOVEMENTS

A second type of cellular motion, *ciliary movement*, is i whiplike movement of cilia on the surface of cells. This occurs in only two places in the human body: on he inside surfaces of the respiratory airways and on he inside surfaces of the uterine tubes (fallopian tubes) of the reproductive tract. In the nasal cavity and lower respiratory airways, the whiplike motion of the cilia causes a layer of mucus to move at a rate of about 1 cm/min toward the pharynx, in this way continually clearing these passageways of the mucus and any particles that have become entrapped in the mucus. In the uterine tubes, the cilia cause slow movement of fluid from the ostium of the uterine tube toward the uterine cavity; it is this movement of fluid that transports the ovum from the ovary to the uterus.

As illustrated in Figure 2–17, a cilium has the appearance of a sharp-pointed curved hair that projects 3 to 4 microns from the surface of the cell. Many cilia project from each single cell—for instance, as many as

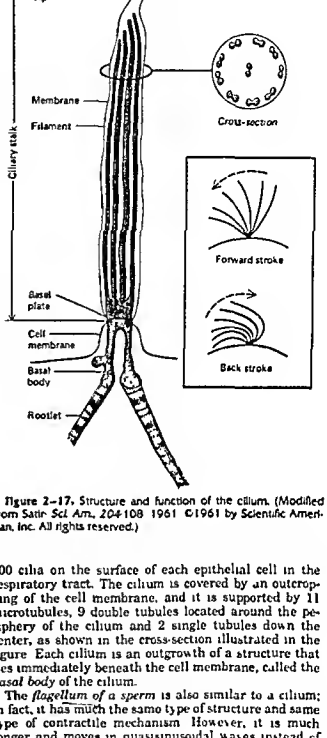


Figure 2–17. Structure and function of the cilium. (Modified from Satir *Sci Am.* 204:108 1961 ©1961 by Scientific American, Inc. All rights reserved.)

200 cilia on the surface of each epithelial cell in the respiratory tract. The cilium is covered by an outcropping of the cell membrane, and it is supported by 11 microtubules, 9 double tubules located around the periphery of the cilium and 2 single tubules down the center, as shown in the cross-section illustrated in the figure. Each cilium is an outgrowth of a structure that lies immediately beneath the cell membrane, called the *basal body* of the cilium.

The *flagellum of a sperm* is also similar to a cilium; in fact, it has much the same type of structure and same type of contractile mechanism. However, it is much longer and moves in quasisinusoidal waves instead of with whiplike movements.

In the inset of Figure 2–17 movement of the cilium is illustrated. The cilium moves forward with a sudden rapid stroke, 10 to 20 times per second, bending sharply where it projects from the surface of the cell. Then it moves backward very slowly in a whiplike manner. The rapid forward movement pushes the fluid lying adjacent to the cell in the direction that the cilium moves, then the slow whiplike movement in the other direction has almost no effect on the fluid. As a result, fluid is continually propelled in the direction of the forward stroke. Since most ciliated cells have large numbers of cilia on their surfaces, and since all the cilia are oriented

in the same direction, this is a very effective means for moving fluids from one part of the surface to another.

Mechanism of Ciliary Movement. Though all aspects of ciliary movement are not yet clear, we do know the following: First, the nine double tubules and the two single tubules are all linked to each other by a complex of protein cross linkages, this total complex of tubules and cross-linkages is called the axoneme. Second, even after removal of the membrane and destruction of other elements of the cilium besides the axoneme, the cilium can still beat under appropriate conditions. Third, there are two necessary conditions for continued beating of the axoneme after removal of the other structures of the cilium: (1) the presence of ATP, and (2) appropriate ionic conditions, including, especially, appropriate concentrations of magnesium and calcium. Fourth, the cilium will continue to beat even after it has been removed from the cell body. Fifth, the tubules on the front edge of the bending cilium slide outward toward the tip of the cilium while the tubules on the back edge of the bending cilium remain in place. Sixth, three protein arms, composed of the protein *dynein* that has ATPase activity, project from each set of peripheral tubules toward the next set.

Given this basic information, it has been postulated that the release of energy from ATP in contact with the ATPase dynein arms causes the arms to "crawl" along the surface of the adjacent pair of tubules. If the front tubules crawl outward while the back tubules remain stationary, this obviously will cause bending.

The way in which cilia contraction is controlled is not understood. However, the cilia of some genetically abnormal cells do not have the two central single tubules and these cilia fail to beat at all. Therefore, it is presumed that some signal, perhaps an electrochemical signal, is transmitted along these two tubules to activate the ATPase activity of the dynein arms, first one of the tubules activating the arms on one side of the cilium followed a fraction of a second later by the other tubule activating the other side, thus causing the forward and backward motion. The cilia usually beat at the frequency of 10 to 20 times per second.

REFERENCES

Agard, D. A. Optical sectioning microscopy: Cellular architecture in three dimensions. *Ann. Rev. Biophys. Bioeng.* 13:191, 1984.

Allen, R. D., and Allen, N. S. Cytoplasmic streaming in amoeboid movement. *Annu. Rev. Biophys. Bioeng.*, 7:469, 1978.

Almers, W., and Stirling, C. Distribution of transport proteins over animal cell membranes. *J. Membr. Biol.* 77:169, 1984.

Baker, P. F., and Knight, D. E. Chemiosmotic hypotheses of exocytosis: A critique. *Biosci. Rep.* 4:285, 1984.

Balaban, R. S. The application of nuclear magnetic resonance to the study of cellular physiology. *Am. J. Physiol.*, 246:C10, 1984.

Blachley, J. E., *et al.* The harmful effects of ethanol on ion transport and cellular respiration. *Am. J. Med. Sci.* 289:22, 1985.

Boyd, A., and Simon, M. Bacterial chemotaxis. *Annu. Rev. Physiol.*, 44:501, 1982.

Bulbring, E. (ed.). Smooth Muscle: An Assessment of Current Knowledge. Austin, University of Texas Press, 1981.

Busch, H. (ed.). The Cell Nucleus: Nuclear Particles. New York, Academic Press, 1981.

Case, R. M. The role of Ca^{2+} stores in secretion. *Cell Calcium*, 1:[illegible], 1981.

Chance, B., *et al.* Hydroperoxide metabolism in mammalian [illegible]. *Physiol. Rev.* 59:527, 1979.

Correia, J. J., and Williams, R. C., Jr. Mechanisms of assembly and disassembly of microtubules. *Annu. Rev. Biophys. Bioeng.*, 12:[illegible], 1983.

Dean, R. T., *et al.* Effects of exogenous amines on mammalian [illegible] with particular reference to membrane flow. *Biochem. J.*, 2[illegible], 1981.

de Petrocirce, M., *et al.* (eds.). Cell Movement and Neoplasia. New York, Pergamon Press, 1980.

DeDuve, C. A Guided Tour of the Living Cell. New York, W. H. Freeman, 1984.

DeDuve, C., and Wattiaux, R. Functions of lysosomes. *Annu. Physiol.* 28:435, 1966.

Dowben, R. M. (ed.). Cell and Muscle Motility. New York, [illegible] Publishing Corp., 1983.

Fawcett, D. W. The Cell. Philadelphia, W. B. Saunders Co., 1981.

Feliger, Y., and Gershon, S. Transport and regulation of [illegible] precursors of mature mitochondrial proteins. *Curr. Top. Cell. [illegible]* 21:217, 1984.

Fine, R. E., and Ochsenfeld, C. D. Supramolecular cytology of vesicles. *Int. Rev. Cytol.* 91:1, 1984.

Fowler, S., and Wolinsky, H. Lysosomes in vascular smooth muscle cells. In Bohr, D. F., *et al.* (eds.), *Handbook of Physiology*, Sec. 2, II. Baltimore, Williams & Wilkins Co., 1980, p. 133.

Frankel, R. B. Magnetic guidance of organisms. *Annu. Rev. Biophys. Bioeng.* 13:85, 1984.

Gerisch, G. Chemotaxis in dictyostelium. *Annu. Rev. Physiol.*, 44:[illegible], 1982.

Gersh, I. (ed.). Submicroscopic Cytochemistry. New York, [illegible] Press, 1973.

Goldman, R. D., *et al.* Cytoplasmic fibers in mammalian cells: Cytoskeletal and contractile elements. *Annu. Rev. Physiol.*, 41:[illegible], 1979.

Goldstein, D. B. The effects of drugs on membrane fluidity. *Annu. Pharmacol. Toxicol.* 24:43, 1984.

Gupta, R. K., *et al.* NMR studies of intracellular metal ions in cells and tissues. *Annu. Rev. Biophys. Bioeng.*, 13:221, 1984.

Hammersen, F. Histology: A Color Atlas of Cytology, Histology, Microscopic Anatomy. Baltimore, Urban & Schwarzenberg, 1980.

Hedi, A. Interactions of water with nonpolar solutes. *Annu. Biophys. Bioeng.* 12:1, 1983.

Korn, E. D. Actin polymerization and its regulation by nonmuscle cells. *Physiol. Rev.* 62:672, 1982.

Kung, C., and Saimi, Y. The physiological basis of taxes in Paramecium. *Annu. Rev. Physiol.* 44:519, 1982.

Macnab, R. M., and Aizawa, S. I. Bacterial motility and flagellar motor. *Annu. Rev. Biophys. Bioeng.*, 13:51, 1984.

Malhotra, S. K. The Plasma Membrane. New York, John Wiley & [illegible], 1983.

Maramorosch, K. (ed.). Advances in Cell Culture. New York, Academic Press, 1981.

McCloskey, M., and Poo, M. M. Protein diffusion in cell membranes: Some biological implications. *Int. Rev. Cytol.*, 87:19, 1984.

Prebble, J. N. Mitochondria, Chloroplasts, and Bacterial Membranes. New York, Longman Inc., 1981.

Prescott, D. (ed.). Methods in Cell Biology: Three-Dimensional Ultrastructure in Biology. New York, Academic Press, 1981.

Schaechter, D. Fluidity and function of hepatocyte plasma membrane. *Hepatology* 4:140, 1984.

Stekhoven, F. S., and Bonting, S. L. Transport adenosine triphosphatases: Properties and functions. *Physiol. Rev.*, 61:1, 1981.

Stephens, R. E., and Edds, K. T. Microtubules: Structure, chemistry, and function. *Physiol. Rev.* 56:709, 1976.

Tseng, H. Atlas of Ultrastructure. New York, Appleton-Century-Crofts, 1968.

Tsien, R. Y. Intracellular measurements of ion activities. *Annu. Rev. Biophys. Bioeng.*, 12:91, 1983.

Wall, D. A., and Maack, T. Endocytic uptake, transport, and catabolism of proteins by epithelial cells. *Am. J. Physiol.*, 244:C12, 1983.

Wheatley, D. N. Mini review: On the possible importance of an intracellular circulation. *Life Sci.*, 36:299, 1985.

Willingham, M. C., and Pastan, I. Endocytosis and exocytosis: Current concepts of vesicle traffic in animal cells. *Int. Rev. Cytol.*, 92:51, 1984.

3

Genetic Control of Protein Synthesis, Cell Function, and Cell Reproduction

Almost everyone knows that the genes control heredity from parents to children, but most persons do not realize that the same genes control the reproduction of and the day-by-day function of all cells. The genes control cell function by determining what substances will be synthesized within the cell—what structures, what enzymes, what chemicals.

Figure 3–1 illustrates the general schema of genetic control. Each gene, which is a nucleic acid called *deoxyribonucleic acid (DNA)*, automatically controls the formation of another nucleic acid, *ribonucleic acid (RNA)*, which spreads throughout the cell and controls the formation of a specific protein. Some proteins are *structural proteins* which, in association with various lipids, form the structures of the various organelles that were discussed in the preceding chapter. But by far the majority of the proteins are *enzymes* that catalyze the different chemical reactions in the cells. For instance, enzymes promote all the oxidative reactions that supply energy to the cell, and they promote the synthesis of various chemicals such as lipids, glycogen, adenosine triphosphate, and so on.

For the formation of most cellular proteins there is only one gene pair in each cell—that is, one gene on each chromosome of a chromosomal pair. If all the DNA of the cell functioned as genes there could theoretically be more than 2 million of these. However, since most of the DNA subserves other functions besides constituting genes, it has been estimated that there are probably about 30,000 *essential* genes in the human cell and perhaps many more that are not essential, which means that 30,000 types of essential proteins are formed in different cells, though not all of these in the same cell for reasons that we will discuss later in the chapter.

THE GENES

The genes are contained, large numbers of them attached end on end, in long, double-stranded, helical molecules of *deoxyribonucleic acid (DNA)* having molecular weights measured in the billions. A very short segment of such a molecule is illustrated in Figure 3–2. This molecule is composed of several simple chemical compounds arranged in a regular pattern explained in the following few paragraphs.

The Basic Building Blocks of DNA. Figure 3–3 illustrates the basic chemical compounds involved in the formation of DNA. These include (1) *phosphoric acid*, (2) a sugar called *deoxyribose*, and (3) four nitrogenous *bases* (two purines, *adenine* and *guanine*, and two pyrimidines, *thymine* and *cytosine*). The phosphoric

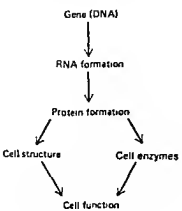


Figure 3–1. General schema by which the genes control cell function.

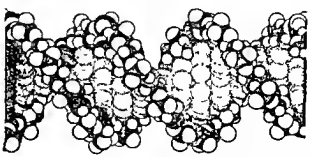


Figure 3–2. The helical, double stranded structure of the gene. The outside strands are composed of phosphoric acid and the sugar deoxyribose. The internal molecules connecting the two strands of the helix are purine and pyrimidine bases, these determine the "code" of the gene.

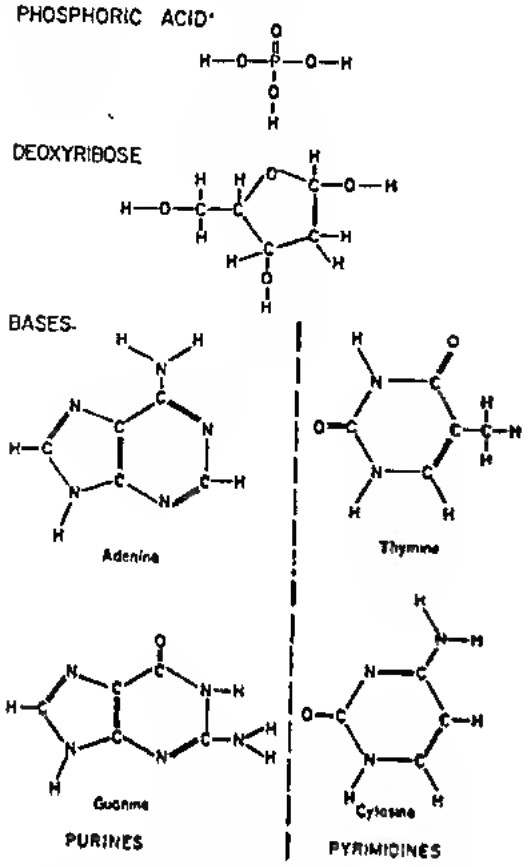


Figure 3–3. The basic building blocks of DNA

acid and deoxyribose form the two helical strands of the DNA molecule, and the bases lie between the strands and connect them together

The Nucleotides. The first stage in the formation of DNA is the combination of one molecule of phosphoric acid, one molecule of deoxyribose, and one of the four bases to form a nucleotide Four separate nucleotides are thus formed, one for each of the four bases, *deoxyadenylic, deoxythymidylic, deoxyguanylic,* and *deoxycytidylic acids* Figure 3–4 illustrates the chemical structure of adenylic acid, and Figure 3–5 illustrates simple symbols for all the four basic nucleotides that form DNA

Organization of the Nucleotides to Form DNA. Figure 3–6 illustrates the manner in which multiple numbers of nucleotides are bound together to form DNA

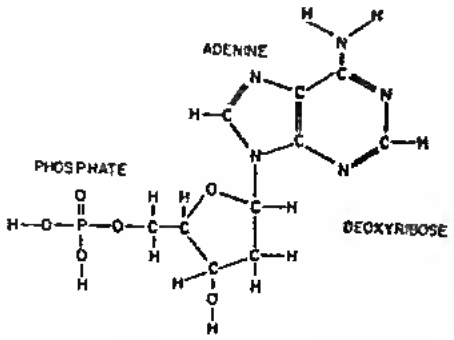


Figure 3–4. Deoxyadenylic acid, one of the nucleotides that make up DNA

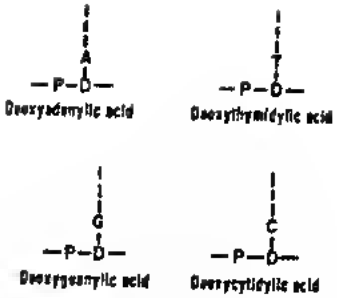


Figure 3–5. Combinations of the basic building blocks o to form nucleotides. (*P*, phosphoric acid, *D*, deoxyribose) four nucleoside bases are *A*, adenine, *T*, thymine; *G*, *C*, cytosine) These four different types of nucleotides DNA

Note that these are combined in such a way phoric acid and deoxyribose alternate with in the two separate strands, and these strands are together by loose bonds between the purine and p Idine bases But note carefully that

(1) the purine base *adenine* always bonds with pyrimidine base *thymine*, and *c* (*realis*)

(2) the purine base *guanine* always bonds with pyrimidine base *cytosine*.

Thus, in Figure 3–6 the sequence of complement pairs of bases is CG, CG, GC, TA, CG, TA, GC, AT, AT However, the bases are bound together by very lo *hydrogen bonding*, represented in the figure by das lines Because of the looseness of these bonds, the t strands can pull apart with ease, and they do so ma times during the course of their function in the cell

Now, to put the DNA of Figure 3–6 into its prop physical perspective, one needs merely to pick up t two ends and twist them into a helix Ten pairs nucleotides are present in each full turn of the helix the DNA molecule, as was illustrated in Figure 3–2

THE GENETIC CODE

The importance of DNA lies in its ability to contr the formation of other substances in the cell. It does th by means of the so-called *genetic code*. When the tw strands of a DNA molecule are split apart, this expos the purine and pyrimidine bases projecting to the sid of each strand It is these projecting bases that form th code

Research studies in the past few years have demon strated that the genetic code consists of successive "trip leta" of bases—that is, each three successive bases a code word. And the successive triplets control th sequence of amino acids in a protein molecule durin its synthesis in the cell. Note in Figure 3–6 that each of the two strands of the DNA molecule carries its own genetic code For instance, the top strand, reading from left to right, has the genetic code GGC, AGA, CTT, the triplets being separated from each other by the arrows As we follow this genetic code through Figures 3–7 and 3–8, we shall see that these three respective triplets are responsible for placement of the three amino acids *proline, serine,* and *glutamic acid*, in a molecule of protein Furthermore, these three amino acids will be lined up in the protein molecule in exactly the same sequence that the genetic code is lined up in this strand of DNA

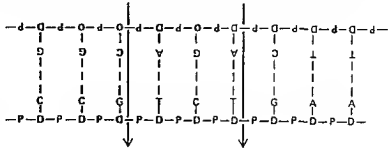


Figure 3–6. Arrangement of deoxyribose nucleotides in DNA.

RIBONUCLEIC ACID (RNA)—THE PROCESS OF TRANSCRIPTION

Since almost all DNA is located in the nucleus of the cell and yet most of the functions of the cell are carried out in the cytoplasm, some means must be available for the genes of the nucleus to control the chemical reactions of the cytoplasm. This is achieved through the intermediary of another type of nucleic acid, *ribonucleic acid (RNA)*, the formation of which is controlled by the DNA of the nucleus; and in the process the code is transferred to the RNA, which is called *transcription*. The RNA is then transported from the nucleus into the cytoplasmic cavity where it controls protein synthesis.

Synthesis of RNA

During synthesis of RNA, the two strands of the DNA molecule separate temporarily; one of these strands is then used as a template for synthesis of the RNA molecules. The code triplets in the DNA cause the formation of *complementary code triplets (called codons)* in the RNA; these codons in turn will control the sequence of amino acids in a protein to be synthesized later in the cytoplasm. When *one strand* of DNA is used in this manner to cause the formation of RNA, the opposite strand remains inactive. Each DNA helix is such a large molecule that it carries the code for more than 1000 genes.

The Basic Building Blocks of RNA. The basic building blocks of RNA are almost the same as those of DNA except for two differences. First, the sugar deoxyribose is not used in the formation of RNA. In its place is another sugar of very slightly different composition, *ribose*. Second, thymine is replaced by another pyrimidine, *uracil*.

Formation of RNA Nucleotides. The basic building blocks of RNA first form nucleotides exactly as described above for the synthesis of DNA. Here again, four separate nucleotides are used in the formation of RNA. These nucleotides contain the bases *adenine, guanine, cytosine,* and *uracil*. Note that these are the same as the bases in DNA except for one of them; the uracil replaces the thymine in DNA.

Activation of the Nucleotides. The next step in the synthesis of RNA is activation of the nucleotides. The occurs by addition to each nucleotide of two phosphate radicals to form triphosphates. These last two phosphates are combined with the nucleotide by *high-energy phosphate bonds* derived from the ATP of the cell.

The result of the activation process is that large quantities of energy are made available to each of the nucleotides, and this energy is used in promoting the subsequent chemical reactions that eventuate in the formation of the RNA chain.

Assembly of the RNA Molecule from Activated Nucleotides Using the DNA Strand as a Template—The Process of Transcription. Assembly of the RNA molecule is accomplished in the manner illustrated in Figure 3–7 under the influence of the enzyme *RNA polymerase*. This is a very large enzyme that has many functional properties necessary for the formation of the RNA molecule. These are:

(1) In the DNA strand immediately ahead of the initial gene is a sequence of nucleotides called the *promoter*. The RNA polymerase has an appropriate complementary structure that recognizes this promoter and be-

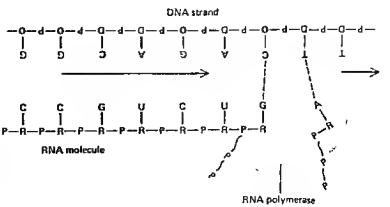


Figure 3–7. Combination of ribose nucleotides with a strand of DNA to form a molecule of ribonucleic acid (RNA) that carries the DNA code from the gene to the cytoplasm. The RNA polymerase moves along the DNA strand and builds the RNA molecule.

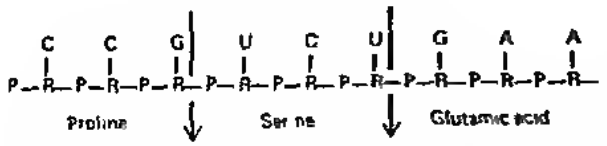


Figure 3–8. Portion of a ribonucleic acid molecule showing three "code" words CCG UCU, and GAA, which represent the three amino acids *proline, serine,* and *glutamic acid*

comes attached to it This is the essential step in initiating the formation of the RNA molecule

(2) Once the RNA polymerase attaches to the promoter, it causes unwinding of about one turn of the DNA helix and separation of the unwound portions of the two strands

(3) Then the polymerase moves along the DNA strand, temporarily unwinding and separating the two DNA strands at each stage of its movement As it moves along, it forms the RNA molecule by the following steps:

(4) First, it causes hydrogen bonds between the bases of the DNA strand and the bases of appropriate RNA nucleotides in the nucleoplasm

(5) Then, one at a time, the RNA polymerase breaks two of the three phosphate radicals away from each of these RNA nucleotides, liberating large amounts of energy from the broken high energy phosphate bonds; this energy is used to cause covalent linkage of the remaining phosphate on the nucleotide with the ribose on the end of the growing RNA molecule

(6) When the RNA polymerase reaches the end of the gene or sequence of genes, it encounters a new sequence of DNA nucleotides called the *chain-terminating sequence;* this causes the polymerase to break away from the DNA strand. Then the polymerase can attach somewhere else to be used again and again for forming new RNA molecules

(7) As the new RNA strand is formed, its hydrogen bonds with the DNA template break away because the complementary strand of DNA has a higher energy of bonding that forces the new RNA strand loose and rebinds the two DNA strands In this way, the RNA molecule is released into the nucleoplasm

It should be remembered that there are four different types of DNA bases and also four different types of RNA nucleotide bases Furthermore, these always combine with each other in specific combinations Therefore, the code that is present in the DNA strand is transmitted in *complementary* form to the RNA molecule. The ribose nucleotide bases always combine with the deoxyribose bases in the following combinations

DNA base	*RNA base*
guanine	cytosine
cytosine	guanine
adenine	uracil
thymine	adenine

After the RNA molecules have been released into the nucleoplasm, they still must be processed further before entering the cytoplasm. The reason for this is that the ly transcribed RNA contains many unwanted se- ences of RNA nucleotides. Some of these occur at both : of the RNA strand, but many others occur in the middle of the strand; this unwanted material comprises, on the average, about five sixths of the total strand Fortunately, a series of enzymes within the nucleoplasm has the capability of cutting out these unwanted se- quences and then resplicing the RNA strand, a proce called *RNA splicing* The RNA is now ready to be use for the formation of proteins.

There are three separate types of RNA, each of whic plays an independent and entirely different role i protein formation These are

(1) *messenger RNA,* which carries the genetic code t the cytoplasm for controlling the formation of the pr teins,

(2) *transfer RNA,* which transports activated acids to the ribosomes to be used in assemb ; protein molecules; and

(3) *ribosomal RNA,* which, along with almost proteins, forms the *ribosomes,* the physical structure on which protein molecules are actually sembled

MESSENGER RNA—THE "CODONS"

Messenger RNA molecules are long strands that suspended in the cytoplasm These molecules composed of several hundred to several thousand n cleotides in unpaired strands, and they contain codo that are exactly complementary to the code triplets the genes Figure 3–8 illustrates a small segment of molecule of messenger RNA Its condons are CCG, UC and GAA These are the codons for the amino ac proline, serine, and glutamic acid Tha transcription these codons from the DNA molecule was demonstrat in Figure 3–7

RNA Codons. Table 3–1 gives the RNA codons f the 20 common amino acids found in protein molecul Note that most of the amino acids are represented more than one codon; also, one codon represents t signal "start manufacturing a protein molecule" a three codons represent "stop manufacturing a prote molecule" In Table 3–1, these two types of codons a designated CI for "chain-initiating" and CT for "chai terminating."

Table 3–1 RNA CODONS FOR THE DIFFERENT AMINO ACIDS AND FOR START AND STOP

Amino Acid	RNA Codons					
Alanine	GCU	GCC	GCA	GCG		
Arginine	CGU	CGC	CGA	CGG	ACA	AC
Asparagine	AAU	AAC				
Aspartic acid	GAU	GAC				
Cysteine	UGU	UGC				
Glutamic acid	GAA	GAG				
Glutamine	CAA	CAG				
Glycine	GGU	GGC	GGA	GGG		
Histidine	CAU	CAC				
Isoleucine	AUU	AUC	AUA			
Leucine	CUU	CUC	CUA	CUG	UUA	UU
Lysine	AAA	AAG				
Methionine	AUG					
Phenylalanine	UUU	UUC				
Proline	CCU	CCC	CCA	CCG		
Serine	UCU	UCC	UCA	UCG	AGC	AG
Threonine	ACU	ACC	ACA	ACG		
Tryptophan	UGG					
Tyrosine	UAU	UAC				
Valine	GUU	GUC	GUA	GUG		
Start (CI)	AUG					
Stop (CT)	UAA	UAG	UGA			

TRANSFER RNA—THE "ANTICODONS"

Another type of RNA that plays a prominent role in protein synthesis is called *transfer RNA* because it transfers amino acid molecules to protein molecules as the protein is synthesized. There are many different types of transfer RNA, but each type will combine specifically with only one of the 20 amino acids that are incorporated into proteins. The transfer RNA then acts as a *carrier* to transport its specific type of amino acid to the ribosomes where protein molecules are formed. In the ribosomes, each specific type of transfer RNA recognizes a particular codon on the messenger RNA, as is described below, and thereby delivers the appropriate amino acid to the appropriate place in the chain of the newly forming protein molecule.

RNA, containing only about 80 nucleotides, is a relatively small molecule in comparison with messenger RNA. It is a folded chain of nucleotides with a cloverleaf appearance similar to that illustrated in Figure 3–9. At one end of the molecule is always an adenylic acid; it is to this that the transported amino acid attaches to a hydroxyl group of the ribose in the adenylic acid. A specific enzyme causes this attachment for each specific type of transfer RNA; this enzyme also determines the type of amino acid that will attach to the respective type of transfer RNA

Since the function of transfer RNA is to cause attachment of a specific amino acid to a forming protein chain, it is essential that each type of transfer RNA also have specificity for a particular codon in the messenger RNA. The specific code in the transfer RNA that allows it to recognize a specific codon is again a triplet of nucleotide bases and is called an *anticodon*. This is located approximately in the middle of the transfer RNA molecule (at the bottom of the cloverleaf configuration illustrated in Figure 3–9). During formation of a protein molecule, the anticodon bases combine loosely by hydrogen bonding with the codon bases of the messenger RNA. In this way the respective amino acids are lined up one after another along the messenger RNA chain, thus establishing the appropriate sequence of amino acids in the protein molecule.

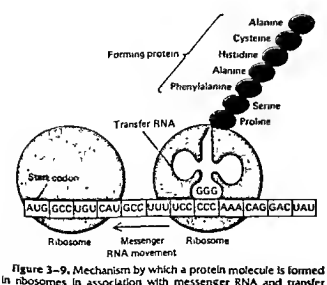


Figure 3–9. Mechanism by which a protein molecule is formed in ribosomes in association with messenger RNA and transfer RNA.

RIBOSOMAL RNA

The third type of RNA in the cell is ribosomal RNA; it constitutes about 60 per cent of the ribosome. The remainder of the ribosome is protein, containing almost 100 different types of proteins that are both structural proteins and enzymes needed in the manufacture of protein molecules.

The ribosome is the physical and chemical structure in the cytoplasm on which protein molecules are actually synthesized. However, it always functions in association with both the other types of RNA as well: transfer RNA transports amino acids to the ribosomes for incorporation into the developing protein molecules, while messenger RNA provides the information necessary for sequencing the amino acids in proper order for each specific type of protein to be manufactured.

The ribosomes are composed of two physical subunits, called the *small subunit*, containing 1 RNA molecule and 33 proteins, and the *large subunit*, containing 3 RNAs and more than 40 proteins. Though we have only partial knowledge of the mechanism of protein manufacture in the ribosome, it is known that messenger RNA and transfer RNA first complex with the small subunit Then the large subunit provides most of the enzymes that promote peptide linkages between the successive amino acids Thus, the ribosome acts as a manufacturing plant in which the protein molecules are formed

Formation of Ribosomes in the Nucleolus. The DNA genes for formation of ribosomal RNA are located in five different chromosomal pairs of the nucleus, and each chromosome contains many duplicates of these genes because of the large amount of ribosomal RNA required for cellular function

As the ribosomal RNA forms, it collects in the *nucleolus*, a specialized structure lying adjacent to the chromosomes. When large amounts of ribosomal RNA are being synthesized, as occurs in cells that manufacture large amounts of protein, the nucleolus is a very large structure, while in cells that synthesize very little protein the nucleolus may not be seen at all. The ribosomal RNA is specially processed in the nucleolus and combined with "ribosomal proteins" to form granular condensation products that are primordial subunits of the ribosomes. These subunits are then released from the nucleolus and transported through the large pores of the nuclear envelope to almost all parts of the cytoplasm Only after the subunits enter the cytoplasm are they assembled to form mature, functional ribosomes. Therefore, proteins are not formed in the nucleus because the nucleus does not contain mature ribosomes.

FORMATION OF PROTEINS ON THE RIBOSOMES—THE PROCESS OF TRANSLATION

When a molecule of messenger RNA comes in contact with a ribosome, it travels along the ribosome, beginning at a predetermined point on the RNA molecule specified by an appropriate sequence of RNA bases. Then, as illustrated in Figure 3–9, while the messenger RNA travels along the ribosome, a protein molecule is formed—a process called *translation*. Thus, the ribosome reads the code of the messenger RNA in much the same way that a tape is "read" as it passes through the

playback head of a tape recorder. Then, when a "stop" (or "chain-terminating") codon slips past the ribosome, the end of a protein molecule is signaled, and the entire molecule is freed into the cytoplasm

Polyribosomes. A single messenger RNA molecule can form protein molecules in several different ribosomes at the same time, the molecule passing along a successive ribosome as it leaves the first, as shown in Figure 3–9 The protein molecules obviously are in different stages of development in each ribosome As a result, clusters of ribosomes frequently occur, three to eight ribosomes being attached together by a single messenger RNA at the same time These clusters are called *polyribosomes*

It is especially important to note that a messenger RNA can cause the formation of a protein molecule in any ribosome, and that there is no specificity of ribosomes for given types of protein The ribosome is simply the structure in which or on which the chemical reactions take place

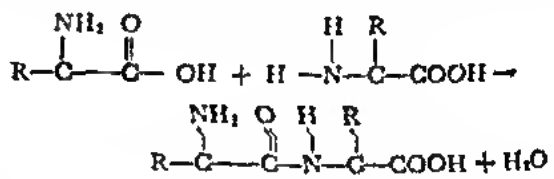
Attachment of Ribosomes to the Endoplasmic Reticulum. In the previous chapter it was noted that many of the ribosomes berome attached to the endoplasmic reticulum This does not occur until the ribosomes begin to form the protein molecules However, the initial ends of certain types of protein molecules immediately attach themselves to specific receptor sites on the endoplasmic reticulum, which causes these molecules to penetrate the reticulum wall and enter the endoplasmic reticulum matrix This occurs while the protein molecule is still being formed by the ribosome, which pulls the ribosome against the endoplasmic reticulum, thereby giving the "granular" appearance to this reticulum The reason these proteins attach to the endoplasmic reticulum is that they have a special sequence of amino acids on their initial ends with specific binding affinity for the receptors on the reticulum These initial amino acids are then removed as the protein enters the reticulum matrix and are replaced by saccharide moieties, giving rise to the typical glycosylated proteins that are processed in the endoplasmic reticulum

Figure 3–10 shows the functional relationship of messenger RNA to the ribosomes and also the manner in which the ribosomes attach to the membrane of the endoplasmic reticulum Note the process of translation occurring in several ribosomes at the same time in response to the same strand of messenger RNA And note also the newly forming polypeptide chains passing through the endoplasmic reticulum membrane into the endoplasmic matrix.

Yet, it should be noted that, except in glandul that form large amounts of protein-rontaining secre vesicles, most proteins formed on the ribosomes released directly into the cytosol and are not of glycosylated variety. These are the enzymes and tural proteins of the cytosol.

Chemical Steps in Protein Synthesis. Some of chemical events that occur in synthesis of a molecule are illustrated in Figure 3–11. This shows representative reactions for three acids, AA_1, AA_2, and AA_n. The stages of the are the following: (1) Each amino acid is *activated* by chemical process in which ATP combines with the acid to form an *adenosine monophosphate compl the amino acid*, giving up two high energy bonds in the process (2) The activated amino having an excess of energy, then *combines with specific transfer RNA to form an amino acid-tRN complex* and, at the same time, releases the monophosphate (3) The transfer RNA carrying amino acid complex then romes in contact with messenger RNA molecule in the ribosome where th anticodon of the transfer RNA attaches temporarily its specific codon of the messenger RNA, thus lining the amino acids in appropriate sequence to form protein molecule. Then, under the influence of the zyme *peptidyl transferase*, one of the proteins in t ribosome, *peptide bonds* are formed between the succe sive amino acids, thus adding progressively to the pr tein chain These chemical events require the energ from two additional high energy phosphate bonds, ma ing a total of four high energy bonds utilized for ea amino acid added to the protein chain. Thus, the sy thesis of proteins is one of the most energy-consumi processes of the cell

Peptide Linkage. The successive amino acids in t protein chain combine with each other according to t following typical reaction.



In this chemical reaction, a hydroxyl radical is remov from the COOH portion of one amino acid while hydrogen of the NH_2 portion of the other amino acid

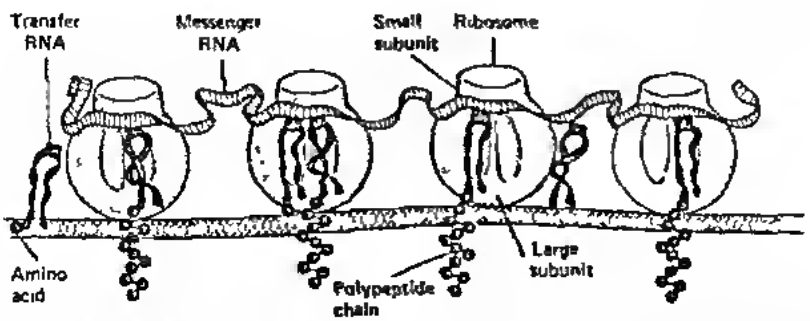
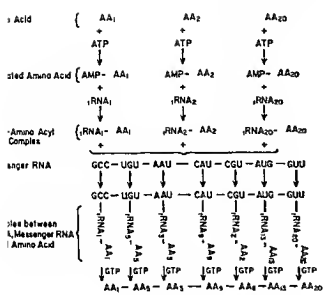


Figure 3–10. An artist's concept of the physical structure of the ribosomes as well as their functional relationship to messenger RNA, transfer RNA, and the endoplasmic reticulum during the formation of protein molecules (From Bloom and Fawcett A Textbook of Histology 10th ed Philadelphia, W B Saunders Company, 1975)



gure 3–11. Chemical events in the formation of a protein ecule.

noved. These combine to form water, and the two ictive sites left on the two successive amino acids nbine, resulting in a single molecule. This process is led *peptide linkage*.

NTHESIS OF OTHER IN THE CELL

Many thousand protein enzymes formed in the man- r just described control essentially all the other chem- il reactions that take place in cells These enzymes omote synthesis of lipids, glycogen, purines, pyrimi- nes, and hundreds of other substances. We will discuss any of these synthetic processes in relation to carbo- drate, lipid, and protein metabolism in Chapters 67 rough 69. It is by means of all these different sub- ances that the many functions of the cells are per- rmed.

ONTROL OF GENETIC FUNCTION ND BIOCHEMICAL ACTIVITY IN ELLS

There are basically two different methods by which ie biochemical activities in the cell are controlled. One these is called *genetic regulation*, in which the activ- ies of the genes themselves are controlled, and the her *enzyme regulation*, in which the activity rates of ie enzymes within the cell are controlled.

Strange as it may seem, every nucleated cell of the ody contains the same set of genes as those originally resent in the fertilized ovum from which the body was rmed. Therefore, the vast differences in the different ypes of cells result not from different types of genes in ie cells but instead from repression or activation of ifferent genes in different cells. This effect is called *ifferentiation* of the cells, a subject that we will discuss i more detail later in the chapter. The respective genes re repressed or activated by special intracellular ge- etic control mechanisms that allow expression of different functional characteristics in separate types of "differentiated" cells, some providing muscle activity, others glandular secretion, and still others the many other bodily functions

Another important function of the intracellular control mechanisms is to control the rates of synthesis of the different intracellular chemicals. It would be unwise for a cell to continue forming biochemical products that are not needed in the cell. Or under other conditions, a cell may need an excess of one particular product but very little of others. For these reasons, a host of intracellular control systems is present in each cell for maintaining appropriate functional balance among the many biochemical synthetic processes

GENETIC REGULATION

The Operon and Its Control of Biochemical Synthesis—Function of the "Promoter." The synthesis of a cellular biochemical product usually requires a series of reactions, and each of these reactions is catalyzed by a special protein enzyme Formation of all the enzymes needed for the synthetic process is often controlled by a sequence of genes located in series one after the other on the same chromosomal DNA strand. This area of the DNA strand is called an *operon*, and the genes responsible for forming the respective enzymes are called the *structural genes* In Figure 3–12, three respective structural genes are illustrated in an operon, and it is shown that they control the formation of three respective enzymes utilized in a particular biochemical synthetic process

Now note in the figure the segment on the DNA strand called the *promoter* This is a series of nucleotides that has specific affinity for RNA polymerase, as already discussed. The polymerase must bind with this promoter before it can begin traveling along the DNA strand to synthesize RNA. Therefore, the promoter is the essential element in activating the operon

Control of the Operon by a "Repressor Protein"—The "Repressor Operator." Also note an additional band of nucleotides lying in the middle of the promoter This area is called a *repressor operator* because a "regulatory" protein can bind here and prevent attachment of RNA polymerase to the promoter, thereby blocking transcription of the genes Such a regulatory protein is called a *repressor protein*. However, each regulatory repressor protein generally exists in two

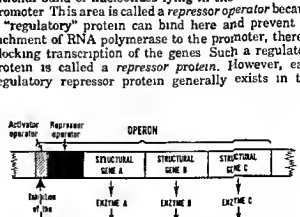


Figure 3–12. Function of the operon to control biosynthesis. Note that the synthesized product exerts a negative feedback to inhibit function of the operon, in this way automatically controlling the concentration of the product itself.

different allosteric forms, one that can bind with the operator and repress transcription and the other that does not bind. That is, one of its forms might be a straight protein molecule and the other the same molecule bent at an angle in the middle. Only one of these forms can repress the operator In turn, varlaus nonprotein substances in the cell, such as some af the cell metabolites, can bind with the repressor protein to change its allosteric state A substonce that causes it to change so that it will bind with the operator and stop transcription is called a *repressor substance* nr *inhibitor substance* On the other hand, a substance that changes the repressor protein so that it will break its bond with the operator is called an *activator substance* ar *inducer substance* because it activates, or induces, the transcription process by removing the repressor peotein

To illustrate the control of gene transcriptian by a repressor peotein, let us give an example The saccharide lactose usually is not available to cells as a food substrate Therefore, the cells normally will not synthesize the required enzymes for metabolic breakdown of lactose. However, when lactose does become available, it induces an allosteric conformational change in a repressor protein, causing it to break away from a repressor operator on the operon that tronscribes for the necessary metabolic enzymes Therefore the operon becomes derepressed, and within a few hours the appropriate enzymes aro present in the cell to rouse breakdown of the lactose Then, as the lactose begins to disappear from within the cell, the rats of synthesis of the enzymes also decreases back to that level only requlred for the amount of lactose that is available Thus one can see the logic of having such regulatory systems in the cell

Control of the Operon by an "Activator Protein"—The "Activatar Operatar." Now note in Figure 3–12 another operator, called the *activator operotor*, that lies adjacent to but ahead of the promoter. When a regulatory protein binds to this operator, it helps ta attract the RNA polymerase to the promater, in this way activating the operon Therefore, a regulatary protein of this type is called an *activator protein* The aperon can be activated or inhibited through the activator operator in ways exactly opposite to the conteol of the repressor operator

Negative Feedback Control of the Operon. Finally, note in Figure 3–12 that the presence of a critical amount of a synthesized product in the cell can cause negative feedback inhibition of the operon that is responsible for its synthesis. It con da this either by causing a regulatory repressor protein ta bind at the repressor operator or a regulatory activator protein to break its bond with the activator operator In either case, the opecon becomes inhibited Therefore, ance the required synthesized product has become abundant enough, the operon becomes dormant On the ather hand, when the synthesized product becomes degraded in the cell and its concentration falls, the aperon once again becomes active In this way, the concentration of the product is automatically conteolled

Other Mechanisms for Control of Transcription by the Operan. Variations in the basic mechanism for conteol of the operon have been discovered with rapidity in the last few years Without giving details, let us list some of these

(1) An opecon is frequently controlled by a *regulatory gene* located elsewhere in the genetic romplex af the nucleus That is, the regulatory gene causes the formatian af a regulatory peotein that in turn acts either a an activator or a repressor substance to control t aperon.

(2) Occasionally, many different opecons are controll at the same time by the same regulatory protein. In in some instances, the same reg as an activator for one operon and as a repressor f another operon When multiple opecons are controll simultanrously in this manner, all the aperons th functian together are called a *regulon*.

(3) Some operons are rontrolled not at the start point af transcription on the DNA strand but inste farther along the strond And, sometimes, the control not even at the DNA strand itself but is instead the processing of the RNA molecules in the nucle before they are released into the cytoplasm; or, rare conteol might occur at the level of RNA translation the ribosomes

Because there are as many as 30,000 to 100,0 different genes in each cell, the large number of diff ways in which genetic activity can be ronteo surprising Through these control mechanisms, cially through negative feedback controls, the cell able ta maintain appropriate concentrations of its essary functional biochemicals The genetic control tems are especially important for controlling the cellular roncentrations af amino acids, amino derivatives, and many if not most of the substrates of carbohydrate, lipid, and protein lism

CONTROL OF ENZYME ACTIVITY

In the same way that inhibitore and acti affect the genetic regulatory system, so also can enzymes themselves be directly rontrolled by other hibitors ar activators This, therefore, represents a s ond category of mechanisms by which ical functions can be controlled

Enzyme Inhibition. A great many of the substances formed in the cell have a direct feedb effect in inhibiting the respective enzyme systems th synthesize them Almost always the synthesized prod acts on the first enzyme in a sequence, rather than the subsequent enzymes, usually binding directly w the enzyme and causing an allosteric ronformatior change that Inactivates it. One can readily recogn the importance of inactivating this first enzyme prevents buildup of intermediary products that will be utilized.

This process of enzyme inhibition is another exam of negative feedback control, it is responsible for c trolling the introcellular conrentrations of some of t amino acids that are not controlled by the gene mechanism as well as the concentrations af the purin the pyrimidines, vitamins, and other substances

Enzyme Activation. Enzymes that are either n mally inactive or that have been Inactivated by so inhibitor substance often can be activated. An exam of this is the action of cyclic adenosine monophosph (AMP) in causing glycogen to split and release gluc malecules to form high energy ATP as discussed in t previous chapter. When most of the adenosine triph phate has been depleted in a cell, a ronsiderable amou of cyclic AMP begins to be formed as a breakdo product af the ATP; the presence of this cyclic A indicates that the cellular reserves of ATP have

ached a low ebb. However, the cyclic AMP immedi-
ely activates the glycogen-splitting enzyme phosphor-
ase, liberating glucose molecules that are rapidly
ed for replenishment of the ATP stores. Thus, in this
se the cyclic AMP acts as an enzyme activator and
ereby helps control intracellular ATP concentration.
Another interesting instance of both enzyme inhibi-
on and enzyme activation occurs in the formation of
e purines and pyrimidines. These substances are
eeded by the cell in approximately equal quantities for
rmation of DNA and RNA. When purines are formed,
ey *inhibit* the enzymes that are required for formation
additional purines. However, they *activate* the en-
mes for formation of the pyrimidines. Conversely, the
rimidines inhibit their own enzymes but activate the
rine enzymes. In this way there is continual cross-
ed between the synthesizing systems for these two
bstances, resulting in almost exactly equal amounts
the two substances in the cells at all times.

In summary, there are two principal methods by which
e cells control proper proportions and proper quan-
ties of different cellular constituents. (1) the mecha-
ism of genetic regulation, and (2) the mechanism of
nzyme regulation. The genes can be either activated or
inhibited, and, likewise, the enzyme systems can be
ither activated or inhibited. Most often, these regula-
ory mechanisms function as feedback rontrol systems
hat continually monitor the cell's biochemical compo-
[illegible] make corrections as needed. But, on occasion,
ubstances from without the cell (especially soma of the
ormones that will be discussed later in this text) also
ontrol the intracellular biochemical reactions by acti-
ating or inhibiting ona or more of the intracellular
ontrol systems.

ELL REPRODUCTION

Cell reproduction is another example of the pervading, ubiquitous role that the DNA-genetic system plays in all life processes. It is the genes and their regulatory mechanisms that determine the growth characteristics of the cells and also when or whether these cells will divide to form new cells. In this way, this all-important genetic system controls each stage in the development of the human being from the single-cell fertilized ovum to the whole functioning body. Thus, if there is any central theme to life it is the DNA-genetic system.

The Life Cycle of the Cell. The life cycle of a cell is the period of time from reproduction to reproduction. When cells are not inhibited and are reproducing as rapidly as they can, this life cycle lasts for 10 to 30 hours. It is terminated by a series of distinct physical events called *mitosis* that cause division of the cell into two new daughter cells. The events of mitosis are illustrated in Figure 3–13 and will be described later. The actual stage of mitosis, however, lasts for only about 30 minutes, so that more than 95 per cent of the life cycle is represented by the interval between mitosis, called *interphase*. In fact, except in special conditions of rapid cellular reproduction, inhibitory controls almost always slow or stop the uninhibited life cycle of the cell. Therefore, different cells of the body have life cycle periods that vary from as little as 10 hours for stimulated bone marrow cells to an entire lifetime of the human body for nerve and striated muscle cells.

Immediately after new daughter cells are formed, they

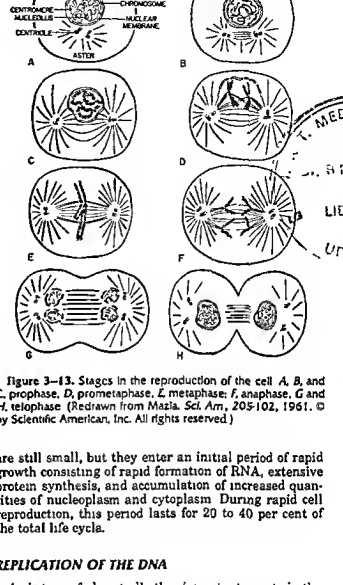


Figure 3–13. Stages in the reproduction of the cell *A, B,* and *C,* prophase. *D,* prometaphase. *E,* metaphase; *F,* anaphase. *G* and *H,* telophase (Redrawn from Mazia. *Sci. Am*, 205:102, 1961. © by Scientific American, Inc. All rights reserved.)

are still small, but they enter an initial period of rapid growth consisting of rapid formation of RNA, extensive protein synthesis, and accumulation of increased quantities of nucleoplasm and cytoplasm During rapid cell reproduction, this period lasts for 20 to 40 per cent of the total life cycle.

REPLICATION OF THE DNA

As is true of almost all other important events in the cell, reproduction begins in the nucleus itself. The first step is *replication (duplication) of all DNA in the chromosomes* Only after this has occurred can mitosis take place

The DNA begins to be duplicated some 5 to 10 hours before mitosis, and this is completed in 4 to 8 hours. The DNA is duplicated only once, so that the net result is two exact *replicates* of all DNA These replicates, in turn, become the DNA in the two new daughter cells that will be formed at mitosis Following replication of the DNA there is another period of 1 to 2 hours before mitosis begins abruptly. However, even during this period, preliminary changes are already beginning to take place that will lead to the mitotic process.

Chemical and Physical Events of DNA Replication. DNA is replicated in much the same way that RNA is transcribed by DNA except for a few important differences:

(1) Both strands of the DNA are replicated, not simply one of them.

(2) Both entire strands of the DNA helix are replicated from end to end rather than small portions of them as occurs in the transcription of RNA by the genes. These two DNA strands contain more than 1000 genes.

(3) The principal enzyme for replicating DNA is *DNA polymerase* rather than RNA polymerase. It lines up the DNA nucleotides in the appropriate order while the enzyme *DNA ligase* causes attachment of the successive nucleotides to each other, utilizing high energy phosphate bonds to energize these attachments.

(4) Formation of each new DNA strand occurs simultaneously in multiple segments at many different points along each of the two strands of the helix. Then the ends of the subunits are joined together by the DNA ligase enzyme.

(5) Each newly formed strand of DNA remains attached to the original DNA strand that is used as its template. Therefore, two new DNA helixes are formed that are exact duplicates of each other and that are still coiled together.

(6) Because the DNA helixes in each chromosome are about 6 cm in length and have millions of turns in each helix, it would be impossible for the two newly formed DNA helixes to uncoil from each other were it not for some special mechanism. This is achieved by enzymes that periodically cut each helix along its entire length, rotate each segment, and then resplice the helix. Thus, the two new helixes become uncoiled.

DNA Repair and DNA "Proofreading." During the few hours between DNA replication and the beginning of mitosis, there is a period of very active repair and "proofreading" of the DNA strands. That is, wherever inappropriate DNA nucleotides have been matched up with the nucleotides of the original template strand, special enzymes cut out the defective areas and replace these with the appropriate complementary nucleotides. This is achieved by the same DNA polymerase and DNA ligase that are used in the process of replication. This repair process is referred to as *DNA proofreading*.

Because of repair and proofreading, the transcription process almost never makes a mistake. But when a mistake is made, this is called a *mutation*, it in turn will cause the formation of some abnormal protein in the cell, often leading to abnormal cellular function and sometimes even to death. Yet, because of the precision of the transcription process, it has been calculated that each human gene mutates only about once in every 200,000 years of human life. Nevertheless, when one realizes that there are 30,000 to 100,000 genes in the human genome and that the period from one generation to another is about 30 years, one would still expect as many as three to ten mutations in the passage of the genome from parent to child. Fortunately, though, each human genome is represented by two separate sets of chromosomes with almost identical genes so that one functional gene almost always is still available to the child despite a mutation.

THE CHROMOSOMES AND THEIR REPLICATION

The DNA helixes of the nucleus are packaged in the chromosomes. The human cell contains 46 chromosomes arranged in 23 pairs. Most of the genes in the two chromosomes of each pair are identical or almost identical with each other, so that it is usually stated that the different genes also exist in pairs, though occasionally this is not the case.

Aside from the DNA in the chromosome there is also a large amount of protein, composed mainly of many small molecules of electropositively charged *histones*. The histones are organized into vast numbers of bobbin-like cores. Successive segments of each DN helix are coiled sequentially around one core after a other. Then the successive cores are packed one on t of the other, thus allowing the tremendously long DN molecule, having a linear length of 6 cm and a molecu weight of about 60 billion, to be packaged in a cell and folded arrangement that is only a few microns length.

The histone cores of the chromosome probably al play an important role in the regulation of DNA actiu because as long as the DNA is packaged tightly it can function as a template either for the formation of RN or for replication of new DNA. Furthermore, some the regulatory proteins have been shown to decondens the histone packaging of the DNA and to allow lo segments at a time to form RNA. Thus, this is sti higher order of regulation than the types of regulat discussed earlier.

Several nonhistone proteins are also major com nents of chromosomes, probably functioning mainly connection with the genetic regulatory machinery, ing as activators, inhibitors, and enzymes.

Replication of the chromosomes follows as a nat result of replication of the DNA helixes, the new D helixes collecting new protein molecules as needed. this stage, the two newly formed chromosomes are ca *chromatids*. They remain temporarily attached to e other (until the time for mitosis) at a point called *centromere* located near the center of each of the c matids.

MITOSIS

The actual process by which the cell splits into new cells is called *mitosis*. Once the genes have b duplicated and each chromosome has split to form two chromatids, mitosis follows automatically withi hour or two.

The Mitotic Apparatus. One of the first event mitosis takes place in the cytoplasm, occurring du the latter part of interphase and early part of proph in or around the small structures called *centrioles* illustrated in Figure 3-13, two pairs of centrioles close to each other near one pole of the nucleus. (Th centrioles, like the DNA and chromosomes, had been replicated during interphase, usually shortly be the replication of the DNA.) Each centriole is a sm cylindrical body about 0.4 micron long and about micron in diameter, consisting mainly of nine para tubule-like structures arranged around the inner of the cylinder. The two centrioles of each pair lie right angles to each other.

Shortly before mitosis is to take place, the two p of centrioles begin to move apart from each other. is caused by protein microtubules growing between respective pairs and actually pushing them apart. the same time, other microtubules grow radially a from each of the centriole pairs, forming a spiny s called the *aster*, in each end of the cell. Some of spines penetrate the nucleus and will play a rol separating the two sets of chromatids during mit The complex of microtubules connecting the two triole pairs is called the *spindle*, and the entire se microtubules plus the two pairs of centrioles is ca the *mitotic apparatus*.

Prophase. The first stage of mitosis, called *prophase*, shown in Figure 3–13A, B, and C. While the spindle forming, the chromosomes of the nucleus, which in :erphase consist of loosely coiled strands, become con-nsed into well-defined chromosomes.

Prometaphase. During this stage (Fig. 3–13D), the clear envelope fragments. At the same time, a new : of microtubules begins to extend outward from a iall condensed portion of each chromatid called the *.etochore*, located on the outside of the centromere ıere the two chromatids are attached to each other ıese new microtubules, in turn, either become attached or interact with the microtubules from the two asters the mitotic apparatus, with one chromatid connecting one of the asters and the other chromatid connecting the opposite aster,

Metaphase. During metaphase (Fig 3–13E), the two ters of the mitotic apparatus are pushed farther apart additional growth of the mitotic spindle Simultane-ıisly the chromatids are pulled tightly by the attached icrotubules to the very center of the cell, lining up to rm the *equatorial plate* of the mitotic spindle

Anaphase. During this phase (Fig 3–13F), the two ıromatids of each chromosome are pulled apart at the ntromere. Exactly how the microtubular system ıuses this is not known, except that it is known that ıe microtubules contain *actin*, which is one of the ntractile proteins of muscle. Therefore, it has been ·esumed that the microtubules might contract or that ıe chromosomal tubules might in some way interact ıth the microtubules of the aster to cause the pulling rce. Regardless of the mechanism, all 46 pairs of ıromatids are separated, forming two separate sets of 3 *daughter chromosomes*. One of these sets is pulled ıward one mitotic aster and the other toward the other ıster at the two respective poles of the dividing cell

Telophase. In telophase (Fig. 3–13G and H), the two ıts of daughter chromosomes are now pulled completely part. Then the mitotic apparatus dissolutes and a new ıuclear membrane develops around each set of chro-ıosomes, this membrane being formed from portions of ıe endoplasmic reticulum that are already present in ıe cytoplasm. Shortly thereafter, the cell pinches in ·vo midway between the two nuclei, for reasons mainly nexplained at present, except. (1) It is the two asters ıat initiate the cell division, each of the two daughter ·lls being composed of the portion of the cell occupied y a dissoluting aster. And (2) a contractile ring of ıicro*filaments* composed of *actin* and probably *myosin*, ıe two contractile proteins of muscle, develops at the ıncture of the newly developing cells and pinches them ff from each other.

·ONTROL OF CELL GROWTH ıND REPRODUCTION

In the normal human body, regulation of cell growth ıd reproduction is mainly a mystery. We know that ·ertain cells grow and reproduce all the time, such as ıe blood-forming cells of the bone marrow, the germinal ıyers of the skin, and the epithelium of the gut How-ver, many other cells, such as smooth muscle cells, do ot reproduce for many years And a few cells, such as ıe neurons and striated muscle cells, do not reproduce uring the entire life of the person.

If there is an insufficiency of some types of cells in ıe body, these will grow and reproduce very rapidly ntil appropriate numbers of them are again available For instance, seven-eighths of the liver can be removed surgically, and the cells of the remaining one-eighth will grow and divide until the liver mass returns almost to normal. The same effect occurs for almost all glandular cells, for cells of the bone marrow, the subcutaneous tissue, the intestinal epithelium, and almost any other tissue except highly differentiated cells, such as nerve and muscle cells

We know very little about the mechanisms that maintain proper numbers of the different types of cells in the body. However, experiments have shown at least three ways in which growth can be controlled. First, growth is often controlled by *growth factors* that come from other parts of the body Some of these circulate in the blood, but others originate in adjacent tissues For instance, the epithelial cells of some glands, such as the pancreas, will fail to grow without a growth factor from the subtying connective tissue of the gland Second, most normal cells will stop growing when they have run out of space for growth This occurs when cells are grown in tissue culture, the cells will grow until they contact a solid object and then growth will stop Third, cells grown in tissue culture often will stop growing when minute amounts of their own secretions are allowed to collect in the culture medium This, too, could provide a means for feedback control of growth

Regulation of Cell Size. Cell size is determined almost entirely by the amount of DNA in the nucleus If replication of the DNA does not occur, the cell grows to a certain size and thereafter remains at that size On the other hand, it is possible, by use of the chemical *colchicine*, to prevent formation of the mitotic spindle and therefore to prevent mitosis also, even though replication of the DNA continues In this event, the nucleus then contains far greater quantities of DNA than normally, and the cell grows proportionately larger It is assumed that this results simply from increased production of RNA and cell proteins, which in turn cause the cell to grow larger

CELL DIFFERENTIATION

A special characteristic of cell growth and cell division is *cell differentiation*, which means changes in physical and functional properties of cells as they proliferate in the embryo to form the different bodily structures However, our problem here is not to describe the stages of differentiation but simply to discuss the theories concerning the means by which the cells change their characteristics to form all the different organs and tissues of the body

The earliest and simplest theory for explaining differentiation was that the genetic composition of the nucleus undergoes changes during successive generations of cells in such a way that one daughter cell inherits a different set of genes from that of the other daughter cell

However, this theory has now been completely disproved in many ways but especially well illustrated by the following simple experiment The nucleus from an intestinal mucosal cell of a frog, when surgically implanted into a frog ovum from which the original nucleus has been removed, will often cause the formation of a completely normal frog. This demonstrates that even the intestinal mucosal cell, which is a reasonably well-differentiated cell, still carries all the nece information for development of all ·ıs in the frog's body.

Therefore, it has become clear that differentiation results not from loss of genes but instead from selective repression of different genetic operons. In fact, electron micrographs suggest that some segments of DNA helixes wound around histone cores become so condensed that they will no longer uncoil to form RNA molecules. One suggestion for the cause of this effect is the following: It has been supposed that a regulatory gene in the genome begins to produce a regulatory protein that causes positive feedback activation of this same regulatory gene, but the same regulatory protein represses another select group of genes. Once the positive feedback process begins, it never stops so that the repressed genes will never function again. Regardless of the mechanism, most mature cells of the human being produce about 8000 to 10,000 proteins rather than the potential 30,000 or more if all genes were active.

Embryological experiments show also that certain cells in an embryo control the differentiation of adjacent cells. For instance, the *primordial chordamesoderm* is called the *primary organizer* of the embryo because it forms a focus around which the rest of the embryo develops. It differentiates into a *mesodermal axis* containing segmentally arranged *somites* and, as a result of *inductions* in the surrounding tissues, causes formation of essentially all the organs of the body.

Another instance of induction occurs when the developing eye vesicles come in contact with the ectoderm of the head and cause it to thicken into a lens plate that folds inward to form the lens of the eye. It is possible that the entire embryo develops as a result of such inductions, one part of the body affecting another part and this part affecting still other parts.

Thus, although our understanding of cell differentiation is still hazy, we know many different control mechanisms by which differentiation *could* occur.

CANCER

Cancer is caused in all or almost all instances by *mutation* of cellular genes that control cell growth and cell mitosis. The mutated genes are called *oncogenes*. Usually two or more different oncogenes must occur in a cell before the cell will become cancerous.

Only a minute fraction of the cells that mutate in the body ever lead to cancer. There are several reasons for this. First, most mutated cells have less survival capability than normal cells and therefore simply die. Second, only a few of the mutated cells that do survive lose the normal feedback controls that prevent excessive growth. And, third, those cells that are potentially cancerous are usually destroyed by the body's immune system before they grow into a cancer. This occurs in the following way: Most mutated cells form abnormal proteins within their cell bodies because of their altered genes, and these proteins then stimulate the body's immune system, causing it to form antibodies against the cancerous cells, in this way destroying them. Indeed, it is believed that all of us are continually forming cells that are potentially cancerous but that our immune system acts as a scavenger that nips these abnormal cells in the bud before they can become established. In support of this is the fact that in persons whose immune systems have been suppressed, such as those who are taking immunosupressant drugs following transplantation of a kidney or a heart, the probability of developing a cancer is multiplied severalfold.

But what is it that causes the mutations themselves? When one realizes that many trillions of new cells are formed each year in the human being, this question should probably better be asked in the following form: Why is it that we do not develop literally millions or billions of mutant cancerous cells? The answer is the incredible precision with which DNA chromosome strands are replicated in each cell before mitosis takes place and also because the "proofreading" process and repairs any abnormal DNA strand before the mitotic process is allowed to proceed. Yet, despite these precautions, probably one newly formed cell in every few million still has significant mutant characteristics.

Thus, chance alone is all that is required for mutations to take place, so we may suppose that a very large number of cancers are merely the result of an unlucky occurrence.

Yet, the probability of mutations can be increased manyfold when a person is exposed to certain chemical, physical, or biological factors. Some of these are the following:

(1) It is well known that *ionizing radiation* such as x rays, gamma rays, and particle radiations from radioactive substances, and even ultraviolet light, can predispose to cancer. Ions formed in tissue cells under the influence of such radiation are highly reactive and rupture DNA strands, thus causing many mutations.

(2) *Chemical substances* of certain types also have a high propensity for causing mutations. Historically, it was long ago discovered that various aniline dye derivatives are very likely to cause cancer, so that workers in chemical plants producing such substances, if unprotected, have a special predisposition to cancer. Chemical substances that can cause mutation are called *carcinogens*. The carcinogens that cause by far the greatest number of deaths in our present-day society are those in cigarette smoke. These cause about one quarter of all cancer deaths.

(3) *Physical irritants* can also lead to cancer, such as continued abrasion of the linings of the intestinal tract by some types of food. The damage to the tissues leads to rapid mitotic replacement of the cells. The more rapid the mitosis, the greater the chance for mutation.

(4) In many families there is a strong *hereditary tendency to cancer*. This probably results from the fact that most cancers require not one mutation but two or more mutations before cancer will occur. In those families that are particularly predisposed to cancer, it is presumed that one or more of the genes are already mutated in the inherited genome. Therefore, far fewer additional mutations must take place in such a person before a cancer will begin to grow.

(5) In experimental animals certain types of viruses can cause some kinds of cancer, including leukemia. This occasionally results by either of two ways. First, in the case of DNA viruses, the DNA strand of the virus can insert itself directly into one of the chromosomes and thereby cause the mutation that leads to cancer. In the case of RNA viruses, some of these carry with them an enzyme called *reverse transcriptase* that will cause DNA to be transcribed from the RNA. Then the transcribed DNA inserts itself into the animal cell chromosome, thus leading to cancer. Yet, despite the demonstration that viral cancer sometimes occurs in animals, it has not yet been proved that cancer spreads this way in human beings, nor that cancer is contagious from one person to another.

Invasive Characteristic of the Cancer Cell. The
o major differences between the cancer cell and the
rmal cell are: (1) The cancer cell does not respect
ual cellular growth limits; the reason for this is that
e cells presumably do not require the growth factors
hat are necessary to cause growth of normal cells. (2)
ancer cells are far less adhesive to each other than are
rmal cells. Therefore, they have a tendency to wander
hrough the tissues, to enter the blood stream, and to
e transported all through the body where they form
adi for numerous new cancerous growths.

Why Do Cancer Cells Kill? The answer to this is
ery simple: Cancer tissue competes with normal tissues
r nutrients. Because cancer cells continue to prolifer-
te indefinitely, their number multiplying day by day,
ne can readily understand that the cancer cells will
oon demand essentially all the nutrition available to
e body. As a result the normal tissues gradually suffer
utritive death.

EFERENCES

ntoniades, H N , and Owen, A J Growth factors and regulation of
cell growth *Annu Rev Med* , *33* 445, 1982
kinson, D E , and Fox, C F (eds) Modulation of Protein Function
New York, Academic Press, 1979
qaq, M , and Blundell, T Evolution and the tertiary structure of
proteins *Annu Rev Biophys Bioeng* , *13* 453 1984
eers, R F (ed) Cell Fusion Gene Transfer and Transformation New
York, Raven Press, 1984
owner, M . Differentiation of Cells New York, Methuen, Inc , 1985
rgant, P J , and Simpson, P Intrinsic and extrinsic control of growth
in developing organs. *Q Rev Biol* , *59* 387, 1984
usch, H (ed). The Cell Nucleus Nuclear Particles New York, Aca-
demic Press, 1981
utler, J G , and Klug, A ' The assembly of a virus *Sci Am* *239*(5) 62
f 1978
lark, B F. C, *et al* (eds) Gene Expression Protein Synthesis and
Control, RNA Synthesis and Control, Chromatin Structure and Func-
tion New York, Pergamon Press, 1978
ohen, S N The manipulation of genes. *Sci Am* , *233*(1) 24, 1975
ohn, W E (ed) Progress in Nucleic Acid Research and Molecular
Biology DNA. Multiprotein Interactions New York, Academic Press,
1981
avidson, R ' Somatic Cell Genetics New York, Van Nostrand Reinhold,
1984
avison, J . Mechanism of control of DNA replication and incompatibil-
ity in ColE1-type plasmids—a review *Gene*, 28 1, 1984
ennisten, K. J (ed) Recombinant DNA New York, Van Nostrand
Reinhold, 1981
eRecondo, A. M (ed) New Approaches in Eukaryotic DNA Replica-
tion New York, Plenum Publishing Corp, 1983
ickerson, R. E Chemical evolution and the origin of life *Sci Am* ,
239 70, 1978
rankel, E ' DNA, The Ladder of Life New York, McGraw Hill, 1978
ilo, N ' Theoretical studies of protein folding *Annu Rev Biophys*
Bioeng , *12* 163, 1983
ail, A. Oncogenes—implications for human cancer A review *J R*
Soc Med , *77* 410, 1984
orton, J D (ed) Development and Differentiation of Vertebrates
New York, Elsevier/North Holland, 1980
Hunt, T (ed) DNA Makes RNA Makes Protein New York Elsevier
Science Publishing Co 1983
Kastrup, K W, and Neilsen, J H (eds) Growth Factors Cellular
Growth Processes, Growth Factors, Hormonal Control of Growth New
York Pergamon Press, 1978
Klevers, R. R , *et al* Cellular clocks and oscillators *Int Rev. Cytol*
86 97, 1984
Kornberg, A DNA Replication San Francisco, W H Freeman, 1980
Kucherlapati, R., and Skoultchi, A I Introduction of purified genes
into mammalian cells *CRC Crit Rev Biochem* , *16* 349 1984
Kulaev, J S (ed) Environmental Regulation of Microbial Metabolism
New York, Academic Press, 1983
Lapis, K (ed) Regulation and Control of Cell Proliferation Heyden,
1984
Leeman, L S , *et al* Sequence-determined DNA separations *Annu*
Rev Biophys Bioeng , *13* 399, 1984
Margalit, H , *et al* Initiation of DNA replication in bacteria Analysis
of an autorepressor control model *J Theor Biol* *111* 183, 1984
Movat, H Z (ed) Inflammatory Reaction New York, Springer Verlag,
1979
Newton C M Biomathematics in oncology Modeling of cellular sys-
tems *Annu Rev Biophys Bioeng* 9 541, 1980
Nicolson, G L Cancer Metastasis *Sci Am* , *240* 66, 1979
Nomura, M *et al* Regulation of the synthesis of ribosomes and
ribosomal components *Annu Rev Biochem* *53* 75 1984
Nordstrom, K , *et al* Control of replication of bacterial plasmids
Genetics, molecular biology, and physiology of the plasmid R1 system
Plasmid, *12* 71, 1984
Ohlendorf, D H , and Matthews B W Structural studies of protein-
nucleic acid interactions *Annu Rev Biophys Bioeng* , *12* 259 1983
Orcutt, B C , *et al* Protein and nucleic acid sequence database systems
Annu Rev Biophys Bioeng *12* 419 1983
Pabo C O and Sauer R T Protein DNA recognition *Annu Rev*
Biochem *53* 293, 1984
Potter, M (ed) Oncogenes in B-Cell Neoplasia Workshop at the
National Cancer Institute National Institutes of Health, Bethesda
MD, March 5–7 1984 New York Springer Verlag 1984
Raff, E C Genetics of microtubule systems *J Cell Biol* 99 1 1984
Rigler R and Wintermeyer W Dynamics of tRNA *Annu Rev*
Biophys Bioeng , *12* 475 1983
Russell T R , *et al* (eds) From Gene to Protein Information Transfer
in Normal and Abnormal Cells New York, Academic Press, 1979
Schimke R T Gene amplification in cultured animal cells. *Cell*, *37* 705,
1984
Schimmel P R , and Redfield A G Transfer RNA in solution Selected
topics *Annu Rev Biophys Bioeng* 9 181, 1980
Schopf, J W The evolution of the earliest cells Sci *Am* *239* 110,
1978
Science Press *Role of RNA in Development and Reproduction* New
York Van Nostrand Reinhold, 1981
Sheffield J B (ed) Cellular Communication During Ocular Develop-
ment New York, Springer-Verlag, 1982
Smith, H S , *et al* The biology of breast cancer at the cellular level
Biochem Biophys Acta, *738* 103 1984
Stoffler, G , and Stoffler-Meilicke, M Immunoelectron microscopy of
ribosomes *Annu Rev Biophys Bioeng* *13* 303 1984
Tanuma, S (Regulation of gene expression and chromatin structure
Role of ADP-ribosylation of chromosomal proteins) *Tanpakushitsu*
Kakusan Koso, *29* 889, 1984
Walker, J E , *et al* The unc operon Nucleotide sequence, regulation
and structure of ATP-synthase *Biochem Biophys Acta*, *768* 164, 1984
Weissman, S M Gene structure and function In Bondy R K , and
Rosenberg, L E (eds) Metabolic Control and Disease, 8th Ed
Philadelphia, W B Saunders Co 1980, p 1
Womble D D , *et al* Analysis of the individual regulatory components
of the IncFII plasmid replication control system *J Bacteriol* , *161* 534,
1985
Wu, R (ed) Recombinant DNA New York, Academic Press, 1979

II

BLOOD CELLS, IMMUNITY, AND BLOOD CLOTTING

4 ■ Red Blood Cells, Anemia, and Polycythemia

5 ■ Resistance of the Body to Infection—The Leukocytes, the Macrophage System, and Inflammation

6 ■ Immunity and Allergy

7 ■ Blood Groups; Transfusion; Tissue and Organ Transplantation

8 ■ Hemostasis and Blood Coagulation

4

Red Blood Cells, Anemia, and Polycythemia

With this chapter we begin a discussion of the blood cells and of other cells closely related to those of the blood the cells of the macrophage system and of the lymphatic system We will first present the functions of red blood cells, which are the most abundant of all the cells of the body and are necessary for delivery of oxygen to the tissues

THE RED BLOOD CELLS

The major function of red blood cells is to transport hemoglobin, which in turn carries oxygen from the lungs to the tissues In some lower animals hemoglobin circulates as free protein in the plasma, not enclosed in red blood cells However, when it is free in the plasma of the human being, approximately 3 per cent of it leaks through the capillary membrane into the tissue spaces or through the glomerular membrane of the kidney into Bowman's capsule each time the blood passes through the capillaries Therefore, for hemoglobin to remain in the blood stream, it must exist inside red blood cells

The red blood cells have other functions besides simply transport of hemoglobin For instance, they contain a large quantity of carbonic anhydrase, which catalyzes the reaction between carbon dioxide and water, increasing the rate of this reaction many thousandfold The rapidity of this reaction makes it possible for the water in blood to react with large quantities of carbon dioxide and thereby transport it from the tissues to the lungs in the form of the bicarbonate ion (HCO_3^-) Also, the hemoglobin in the cells is an excellent acid-base buffer (as is true of most proteins), so that the red blood cells are responsible for as much as 60 per cent of all the buffering power of whole blood These specific functions are discussed elsewhere in the text and, consequently, are merely mentioned at the present time

The Shape and Size of Red Blood Cells. Normal red blood cells are biconcave discs having a mean diameter of approximately 8 microns and a thickness at the thickest point of 2 microns and in the center of 1 micron or less The average volume of the red blood cell is 83 cubic microns

The shapes of red blood cells can change remarkably as the cells pass through capillaries Actually, the red blood cell is a "bag" that can be deformed into almost any shape. Furthermore, because the normal cell has a great excess of cell membrane for the quantity of m terial inside, deformation does not stretch the me brane, and consequently does not rupture the cell, would be the case with many other cells

Concentration of Red Blood Cells in the Blo In normal men the average number of red blood c per cubic millimeter is 5,200,000 (±300,000) and normal women 4,700,000 (±300,000) The number red blood cells varies in the two sexes and at differ ages, as shown in Figure 4–1 Also, the altitude at wh the person lives affects the number of red blood c this is discussed later

Quantity of Hemoglobin in the Cells. Red b cells have the ability to concentrate hemoglobin in t cell fluid up to approximately 34 grams per deciliter cells The concentration never rises above this value this is a metabolic limit of the cell's hemoglobin-form mechanism Furthermore, in normal persons the p centage of hemoglobin is almost always near the m mum in each cell However, when hemoglobin forma is deficient in the bone marrow, the percentage of he globin in the cells may fall considerably below t value and the volume of the red cell may decrease well because of diminished hemoglobin to fill the cel

When the hematocrit (the percentage of the blood b is cells—normally 40 to 45 per cent) and the quant of hemoglobin in each respective cell are normal, whole blood of men contains an average of 16 grams hemoglobin per deciliter and of women an average of grams per deciliter As will be discussed in connec with the transport of oxygen in Chapter 41, each gr of pure hemoglobin is capable of combining with proximately 1 39 milliliters of oxygen Therefore, normal man, over 21 milliliters of oxygen can be carr in combination with hemoglobin in each deciliter blood, and in normal woman 19 milliliters of oxygen c be carried

PRODUCTION OF RED BLOOD CELLS

Areas of the Body That Produce Red Blood Ce In the early few weeks of embryonic life, primit nucleated red blood cells are produced in the yolk s During the middle trimester of gestation the liver is main organ for production of red blood cells, and at same time a reasonable quantity of red blood cells also produced by the spleen and lymph nodes Th

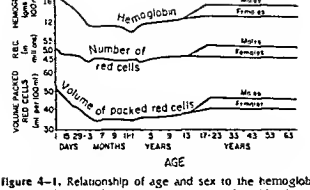


Figure 4–1. Relationship of age and sex to the hemoglobin content, red blood cell count, and hematocrit of the blood

In the latter part of gestation and after birth, red blood cells are produced exclusively by the bone marrow As illustrated by Figure 4–2, the bone marrow of essentially all bones produces red blood cells until a person is 5 years old, but the marrow of the long bones, except for the proximal portions of the humeri and tibiae, becomes quite fatty and produces no more red blood cells after approximately the age of 20 years Beyond this age most red-cells are produced in the marrow of the membranous bones, such as the vertebrae, the sternum, the ribs, and the ilia. Even in these bones the marrow becomes less productive as age increases

Sometimes when various factors stimulate the bone marrow to produce tremendous quantities of red blood cells, much of the marrow that has already stopped producing red blood cells can once again become productive, and marrow that is still producing red blood cells becomes greatly hyperplastic and produces far greater than normal quantities.

Genesis of the Red Blood Cell. In the bone marrow are many *pluripotent hemopoietic stem cells* from which any type of blood cell can be formed. These cells reproduce throughout the life of the person, though their numbers do diminish with age Still other stem cells have become *unipotent*, that is, capable of forming only a particular type of cell, such as red blood cells or white blood cells. These are frequently called *committed stem cells*. However, these unipotent stem cells cannot be identified from their appearance. Instead, they, like the pluripotent stem cells, have an appearance similar to that of large lymphocytes.

The first cell that can be identified as belonging to the red blood cell series is the *proerythroblast*, illustrated in Figure 4–3. Under appropriate stimulation, large numbers of these cells are formed from the stem cells.

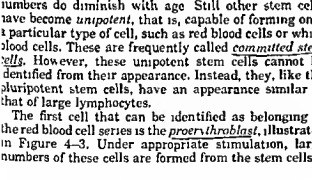


Figure 4–2. Relative rates of red blood cell production in the different bones at different ages

Once the proerythroblast has been formed it divides several more times, eventually forming 8 to 16 mature red blood cells The new cells of this first generation are called *basophil erythroblasts* because they stain with basic dyes; the cell at this time has accumulated very little hemoglobin However, in the next generation of cells, called *polychromatophil erythroblasts*, enough red hemoglobin has begun to be formed to give these cells the polychromatophilic appearance After another division or so, very large quantities of hemoglobin are formed, and the cells then are called *orthochromatic erythroblasts*, which are now colored red by the hemoglobin Finally, after the cytoplasm of these cells has become filled with hemoglobin to a concentration of approximately 34 per cent, the nucleus condenses to a small size and is extruded from the cell. At the same time the endoplasmic reticulum is reabsorbed The cell at this stage is called a *reticulocyte* because it still contains a small amount of basophilic material, consisting of remnants of the Golgi apparatus, mitochondria, and a few other cytoplasmic organelles In this reticulocyte stage the cells pass into the blood capillaries by diapedesis (squeezing through the pores of the membrane)

The remaining basophilic material in the reticulocyte normally disappears within one to two days, and the cell is then the *mature erythrocyte* Because of the short life of the reticulocytes, their concentration among all the red cells of the blood is normally slightly less than 1 per cent

Regulation of Red Blood Cell Production—The Role of Erythropoietin

The total mass of red blood cells in the circulatory system is regulated within very narrow limits, so that an adequate number of red cells is always available to provide sufficient tissue oxygenation and, yet, so that the cells do not become so concentrated that they impede blood flow What we know about this control mechanism is the following

Tissue Oxygenation as the Basic Regulator of Red Blood Cell Production. Any condition that causes the quantity of oxygen transported to the tissues to decrease ordinarily increases the rate of red blood cell production Thus, when a person becomes extremely *anemic* as a result of hemorrhage or any other condition, the bone marrow immediately begins to produce large quantities of red blood cells Also, destruction of major portions of the bone marrow by any means, especially x-ray therapy, causes hyperplasia of the remaining bone marrow, thereby attempting to supply the demand for red blood cells in the body.

At very *high altitudes*, where the quantity of oxygen in the air is greatly decreased, insufficient oxygen is transported to the tissues, and red cells are produced so rapidly that their number in the blood is considerably increased

Therefore, it is obvious that it is not the concentration of red blood cells in the blood that controls the rate of red cell production, but instead it is the functional ability of the cells to transport oxygen to the tissues in relation to the tissue demand for oxygen.

Various diseases of the circulation that cause decreased blood flow through the peripheral vessels, and particularly those that cause failure of oxygen absorption by the blood as it passes through the lungs, can

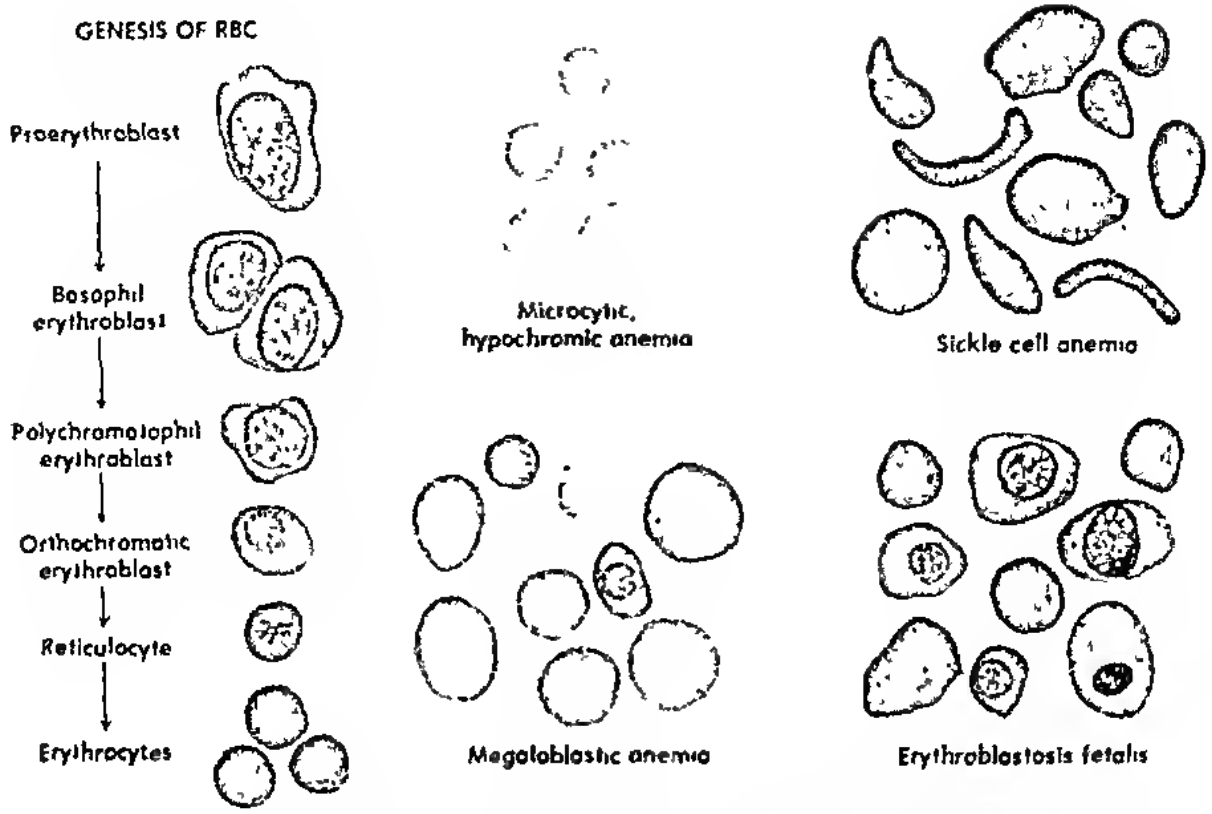


figure 4–3. Genesis of red blood cells and red blood cells in different types of anemias.

also increase the rate of red cell production This is especially apparent in prolonged *cardiac failure* and in many *lung diseases*, for the tissue hypoxia resulting from these conditions increases the rate of red cell production, with resultant increase in the hematocrit and usually increase in the total blood volume as well

Erythropoietin, Its Formation in Response to Hypoxia, and Its Function in Regulating Red Blood Cell Production The principal factor that stimulates red blood cell production is a circulating hormone called *erythropoietin*, a glycoprotein with a molecular weight of about 40,000 In the absence of erythropoietin, hypoxia has either no effect or very little effect in stimulating red blood cell production On the other hand, when the erythropoietin system is functional, hypoxia causes marked increase in erythropoietin production, and the erythropoietin in turn enhances red blood cell production until the hypoxia is relieved

Role of the Kidneys in the Formation of Erythropoietin. In the normal person, 90 to 95 per cent of all erythropoietin is formed in the kidneys However, it is not known exactly where in the kidney it is formed The amount that can be extracted from the kidney cortex is considerably greater than that extracted from the medulla Some research has suggested that erythropoietin might be formed by the *juxtaglomerular cells*, cells in the walls of the afferent arterioles adjacent to the glomeruli Regardless of where erythropoietin is formed in the kidney, not only does hypoxia increase its production but diminished blood flow to the kidney at times also causes excess formation of erythropoietin and therefore leads to excess red blood cell production And, finally, some tumors of the kidney occasionally form large amounts of erythropoietin, this too leading to excess numbers of red blood cells

When both kidneys are removed from a person or when the kidneys are destroyed by renal disease, the person invariably becomes very anemic because th 10 per cent of erythropoietin formed in other ti (mainly in the liver but perhaps also in macroph is sufficient to cause only one third to one half as red blood cell formation as needed by the body.

Effect of Erythropoietin on Erythrogenesis. placing an animal or person in an atmosphere o oxygen, erythropoietin begins to be formed within utes or hours, and it reaches maximum produ within 24 hours Yet, almost no new red blood appear in the circulating blood until about five later From this fact, as well as still other studi has been determined that the important effect of e ropoietin is to stimulate the production of proery blasts from hemopoietic stem cells in the bone ma In addition, once the proerythroblasts are forme erythropoietin causes these cells also to pass mor idly through the different erythroblastic stages normally, also speeding up the production of new The rapid production of cells continues as long a person remains in the low oxygen state or until er red blood cells are produced to carry adequate am of oxygen to the tissues despite the low oxygen; a time the rate of erythropoietin production decrea a level that will maintain the required number cells, but not an excess

Upon removal of a person from a state of low ox the rate of oxygen transport to the tissues rises normal, which causes the rate of erythropoietin f tion to decrease to zero almost instantaneously an rate of red blood cell production to fall essentia zero within several days. Red cell production re at this extremely low level until enough cells have out their life spans and degenerated, so that the t again receive only their normal complement of ox

In the complete absence of erythropoietin, fe blood cells are formed by the bone marrow. At the

treme, when extreme quantities of erythropoietin are med, the rate of red blood cell production can rise to high as ten or more times normal. Therefore, the ythropoietin control mechanism for red blood cell oduction is an extremely powerful one.

System Diagram for Control of Tissue Oxygena-on. Figure 4–4 illustrates a control diagram for the gative feedback mechanism that regulates the level tissue oxygenation. It lists the factors that frequently use tissue hypoxia, which in turn stimulates erythpoietin production in the kidneys. The erythropoietin en enhances the production of red blood cells. If the itiating condition is not too severe, the increased mbers of red blood cells will return the level of tissue ygenation to normal. Thus, this is a typical negative edback control system, illustrating the principles of meostatic control that were discussed in Chapter 1

Effect of Rate of Red Blood Cell Formation on e Type of Cell Released into the Blood. When the ne marrow produces red blood cells at a very rapid te, many of the cells are released into the blood before ey are mature erythrocytes. Thus, very rapid cell oduction can cause the percentage of reticulocytes in e circulating blood to rise from less than 1 per cent to high as 30 to 50 per cent of the total number of red ood cells. If the rate of production is even greater, a rge number of erythroblasts also may appear in the rculating blood. In some severe anemias, such as ythroblastosis fetalis, which is discussed later in the apter and in greater detail in Chapter 7, the concenation of erythroblasta occasionally rises to as high as to 20 per cent of all the circulating red blood cells.

itamins Needed for Formation f Red Blood Cells

The Maturation Factor—Vitamin B_{12} (Cyanocoalamin). Vitamin B_{12} is an essential nutrient for all lls of the body, and growth of tissues in general is reatly depressed when this vitamin is lacking. This sults from the fact that vitamin B_{12} is required for ynthesis of DNA. Therefore, lack of this vitamin causes ilure of nuclear maturation and division. Since tissues hat produce red blood cells are among the most rapidly growing and proliferating of all the body's tissues, lack of vitamin B_{12} especially inhibits the rate of red blood cell production. Furthermore, the erythroblastic cells of the bone marrow, in addition to failing to proliferate rapidly, become larger than normal, developing into so-called *megaloblasts*, and the adult erythrocyte, called a *macrocyte*, has a flimsy membrane and is often irregular, large, and oval instead of the usual biconcave disk These poorly formed macrocytes, after entering the circulating blood, are capable of carrying oxygen normally but their fragility causes them to have a short life, one half to one third of normal Therefore, it is said that vitamin B_{12} deficiency causes *maturation failure* in the process of erythropoiesis

The cause of the abnormally shaped cells seems to be the following The inability of the cells to synthesize adequate quantities of DNA causes slow reproduction of the cells but does not prevent formation of RNA within the cytoplasm Therefore, the quantity of RNA becomes much greater than normal, leading to excess production of all other cytoplasmic constituents as well, including excess hemoglobin, which causes the enlarged cells In addition, some of the genes of the cells are probably replicated abnormally, leading to many other abnormalities of the cells as well

Maturation Failure Caused by Poor Absorption of Vitamin B_{12}—Pernicious Anemia. The most common cause of maturation failure is not lack of vitamin B_{12} in the diet but instead failure to absorb vitamin B_{12} from the gastrointestinal tract This often occurs in the disease called *pernicious anemia*, in which the basic abnormality is an *atrophic gastric mucosa* that fails to secrete normal gastric secretions The parietal cells of the gastric glands secrete a glycoprotein called *intrinsic factor*, which combines with vitamin B_{12} of the food and makes the B_{12} available for absorption by the gut It does this in the following way First, the intrinsic factor binds tightly with the vitamin B_{12} In this bound state the B_{12} is protected from digestion by the gastrointestinal enzymes Second, still in the bound state, the intrinsic factor binds to specific receptor sites on the brush border membranes of the mucosal cells in the ileum Third, vitamin B_{12} is transported into the blood during the next few hours, probably by the process of pinocytosis carrying the intrinsic factor and the vitamin together through the membrane

Lack of intrinsic factor, therefore, causes loss of much of the vitamin because of both enzyme action in the gut and failure of its absorption

Once vitamin B_{12} has been absorbed from the gastrointestinal tract it is stored in large quantities in the liver and then released slowly as needed to the bone marrow and other tissues of the body The total amount of vitamin B_{12} required each day to maintain normal red cell maturation is only about 5 micrograms, and the normal store in the liver and other body tissues is about 1000 times this amount. Therefore, as many as 4 to 5 years of defective B_{12} absorption are required to cause maturation failure anemia

Effect of Folic Acid (Pteroylglutamic Acid) on Red Cell Maturation. Occasionally a patient with maturation failure anemia responds to folic acid therapy instead of to vitamin B_{12}, so it is apparent that this vitamin is also concerned with the maturation of red blood cells. Folic acid, like B_{12}, is required for formation of DNA but in a different way. It promotes the methylation of deoxyuridylate to form deoxythymidylate, one of the nucleotides required for DNA synthesis.

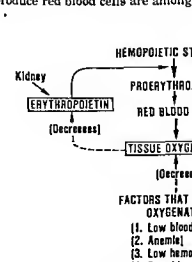


Figure 4–4. Function of the erythropoietin mechanism to increase the production of red blood cells when extraneous factors decrease tissue oxygenation.

FORMATION OF HEMOGLOBIN

Synthesis of hemoglobin begins in the erythroblasts and continues slightly even into the reticulocyte stage, for when the reticulocytes leave the bone marrow and pass into the blood stream they continue to form minute quantities of hemoglobin for another day or so

Figure 4–5 shows the basic chemical steps in the formation of hemoglobin From tracer studies with isotopes it is known that the heme portion of hemoglobin is synthesized mainly from acetic acid and glycine and that most of this synthesis occurs in the mitochondria The acetic acid is changed in the Krebs cycle, which will be explained in Chapter 67, into succinyl-CoA, and then two molecules of this combine with two molecules of glycine to form a pyrrole compound In turn, four pyrrole compounds combine to form a protoporphyrin compound One of the protoporphyrins, known as protoporphyrin IX, then combines with iron to form the heme molecule Finally, four heme molecules combine with a very long polypeptide chain synthesized by the ribosomes, forming a subunit of hemoglobin called a *hemoglobin chain* (Fig 4–6). Each of these chains has a molecular weight of about 16,000, and four of them in turn bind together loosely to form the whole hemoglobin molecule

There are slight variations in the subunit hemoglobin chains, depending on the amino acid composition of its polypeptide portion The different types of chains are designated *alpha chains, beta chains, gamma chains,* and so forth The most common form of hemoglobin in the adult human being, *hemoglobin A*, is a combination of *two alpha chains* and *two beta chains*

Since each chain has a heme prosthetic group, there are four separoto iron atoms in each hemoglobin molecule, each of these can bind with 1 molecule of oxygen, making a total of 4 molecules of oxygen (or 8 atoms) that can be transported by each hemoglobin molecule. Hemoglobin has a molecular weight of 64,458.

The nature of the hemoglobin chains determines the binding affinity of the hemoglobin for oxygen Abnormalities of the chains can alter the physical characteristics of the hemoglobin molecule as well For instance, in sickle cell anemia the amino acid valine is substituted for glutamic acid at one point in each of the two beta chains When this type of hemoglobin is exposed to low oxygen, it forms elongated crystals inside the red blood cells that are sometimes 15 microns in length These make it almost impossible for the cells to pass through the small capillaries, and the spiked ends of the crystals are very likely to rupture the cell membranes, thus leading to sickle cell anemia.

Combination of Hemoglobin with Oxygen. The most important feature of the hemoglobin molecule is

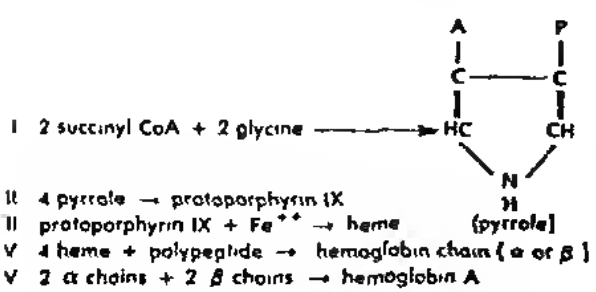


Figure 4–5. Formation of hemoglobin.

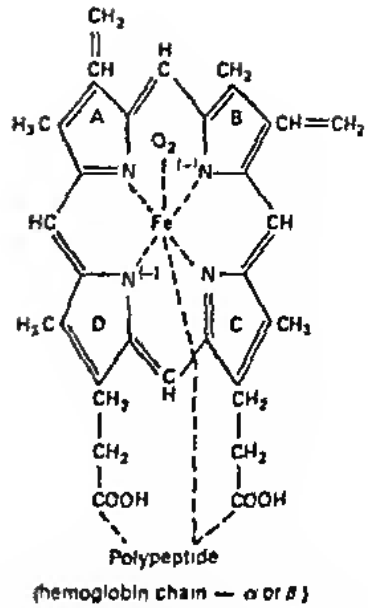


Figure 4–6. Basic structure of the hemoglobin molecule. ing one of the four heme complexes bound with the c globin core of the hemoglobin molecule

its ability to combine loosely and reversibly with oxy This ability will be discussed in detail in Chapter relation to respiration, for the primary function of h globin in the body depends upon its ability to com with oxygen in the lungs and then to release this ox reodily in the tissue capillaries where the gaseous sion of oxygen is much lower than in the lungs

Oxygen *does not* combine with tha two positive ences of the ferrous iron in the hemoglobin mole Instead, it binds loosely with one of the six "coor tion" valences of the iron atom. This is an extre loose bond so that the combination is easily rever Furthermore, the oxygen does not become ionic ox but is carried as molecular oxygen to the tissues w because of the loose, readily reversible combinatio is released into the tissue fluids in the form of diss molecular oxygen rather than ionic oxygen.

IRON METABOLISM

Because iron is important for formation of hemogl myoglobin, and other substances such as the chromes, cytochrome oxidase, peroxidase, and cat it is essential to understand the means by which ir utilized in the body

The total quantity of iron in the body averages 4 grams, approximately 65 per cent of which is pr in the form of hemoglobin. About 4 per cent is pr in the form of myoglobin, 1 per cent in the form o various heme compounds that promote intracel oxidation, 0 1 per cent combined with the protein t ferrin in the blood plasma, and 15 to 30 per cent s mainly in the liver in the forms of ferritin and h siderin

Transport and Storage of Iron. Transport, sto and metabolism of iron in the body are illustrat Figure 4–7, and may be explained as follows When is absorbed from the small intestine, it immedi combines with a beta globulin, *apotransferrin*, to *transferrin*, which is then transported in the plasma. The iron is loosely combined with the glo

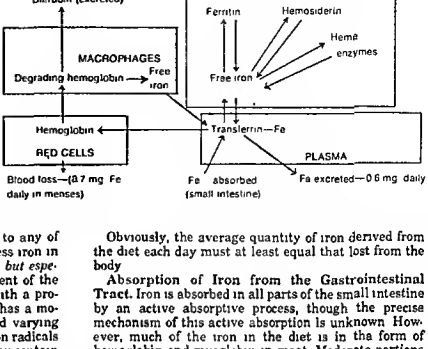


Figure 4–7. Iron transport and metabolism.

molecule and, consequently, can be released to any of the tissue cells at any point in the body Excess iron in the blood is deposited in all cells of the body *but especially in the liver cells*, where about 60 per cent of the excess is stored. There it combines mainly with a protein, *apoferritin*, to form *ferritin* Apoferritin has a molecular weight of approximately 460,000, and varying quantities of iron can combine in clusters of iron radicals with this large molecule; therefore, ferritin may contain only a small amount of iron or a relatively large amount This iron stored in ferritin is called *storage iron*

Smaller quantities of the iron in the storage pool are stored in an extremely insoluble form called *hemosiderin*. This is especially true when the total quantity of iron in the body is more than the apoferritin storage pool can accommodate. Hemosiderin forms large clusters in the cells and consequently can be stained and observed as large particles in tissue slices by usual histologic techniques. Ferritin can also be stained, but the ferritin particles are so small and dispersed that they usually can be seen only with the electron microscope

When the quantity of iron in the plasma falls very low, iron is removed from ferritin quite easily but less easily from hemosiderin. The iron is then transported by the transferrin in the plasma to the portions of the body where it is needed.

A unique characteristic of the transferrin molecule is that it binds with receptors in the cell membranes of erythroblasts in the bone marrow. Then it, along with its bound iron, is ingested into the erythroblasts by endocytosis. Finally, the transferrin delivers the iron directly to the mitochondria where heme is synthesized. In persons who do not have adequate quantities of transferrin in their blood, failure to transport iron to the erythroblasts in this manner causes severe hypochromic anemia—that is, decreased numbers of red cells and even those cells containing very little hemoglobin

When red blood cells have lived their life span and are destroyed, the hemoglobin released from the cells is ingested by the cells of the macrophage-monocyte system. There free iron is liberated, and it can then either be stored in the ferritin pool or be reused for formation of hemoglobin.

Daily Loss of Iron. About 1 milligram of iron is excreted each day by men, mainly into the feces Additional quantities of iron are lost whenever bleeding occurs. Thus, in women, the menstrual loss of blood brings the average iron loss to a value of approximately 2 milligrams per day.

Obviously, the average quantity of iron derived from the diet each day must at least equal that lost from the body

Absorption of Iron from the Gastrointestinal Tract. Iron is absorbed in all parts of the small intestine by an active absorptive process, though the precise mechanism of this active absorption is unknown However, much of the iron in the diet is in the form of hemoglobin and myoglobin in meat Moderate portions of these compounds can be absorbed by pinocytosis directly into the intestinal mucosal cells from the intestinal tract Once in these cells the large molecules are digested by the lysosomes, and the iron is released to the transferrin in the blood capillaries of the intestinal villi

The rate of iron absorption is extremely slow, with a maximum rate of only a few milligrams per day This means that when tremendous quantities of iron are present in the food, only small proportions of this will be absorbed On the other hand, if only minute quantities are present, far greater proportions will be absorbed

Regulation of Total Body Iron by Alteration of Rate of Absorption. When the body has become saturated with iron so that essentially all of the apoferritin in the iron storage areas is already combined with iron, the rate of absorption of iron from the intestinal tract becomes greatly decreased On the other hand, when the iron stores have been depleted of iron, the rate of absorption becomes greatly accelerated, to as much as five or more times as great as when the iron stores are saturated Thus, the total body iron is regulated to a great extent by altering the rate of absorption

Feedback Mechanisms by Which Iron Absorption is Regulated. When essentially all the apoferritin in the body has become saturated with iron, it becomes difficult for transferrin to release iron to the tissues As a consequence, the transferrin, which is normally only one third saturated with iron, now becomes almost fully bound with iron so that the transferrin accepts almost no new iron from the mucosal cells. Then, as a final stage of this process, the buildup of excess iron in the mucosal cells themselves depresses active absorption of iron from the intestinal lumen and at the same time slightly enhances the rate of excretion of iron from the mucosa.

Still another factor that may help control iron absorption is the iron transport protein *apotransferrin*, which not only transports iron in the blood but also enters the

mucosal cells and transports iron through the cytosol of these cells as well It is believed that when the liver becomes saturated with iron the liver produces smaller quantities of apotransferrin, thus reducing the iron transport process all the way from the mucosal cells to all other parts of the body

Yet, despite this feedback control system for regulating iron absorption, when a person eats extreme amounts of iron compounds, excess iron does enter the blood and can lead to massive deposition of hemosiderin in the macrophages throughout the body, which at times can be very damaging

DESTRUCTION OF RED BLOOD CELLS

When red blood cells are delivered from the bone marrow into the circulatory system they normally circulate an average of 120 days before being destroyed. Even though mature red cells do not have a nucleus and also have neither mitochondria nor endoplasmic reticulum, nevertheless they do still have cytoplasmic enzymes that are capable of metabolizing glucose by the glycolytic process, thus forming small amounts of ATP. Also, additional small amounts of ATP can be formed utilizing a pathway called the hexosemonophosphate shunt (see Chapter 67) followed by oxidative formation of ATP The ATP in turn serves the red cell in several important ways (1) to maintain pliability of the cell membrane, (2) to maintain membrane transport of ions, (3) to keep the iron of the cell's hemoglobin in the ferrous form rather than the ferric form (which causes the formation of methemoglobin that will not carry oxygen), and (4) to prevent oxidation of the proteins in the red cell However, these metabolic systems of the red cell become progressively less active with time As the cells become older they become progressively more fragile, presumably because their life processes simply wear out

Once the red cell membrane becomes very fragile, the cell may rupture during passage through some tight spot of the circulation Many of the red cells fragment in the spleen where the cells squeeze through the red pulp of the spleen Here the spaces between the structural trabeculae of the pulp are only 3 microns wide, in comparison with the 8 micron diameter of the red cell When the spleen is removed, the number of abnormal cells and old cells circulating in the blood increases considerably

Destruction of Hemoglobin. The hemoglobin released from the cells when they burst is phagocytized almost immediately by macrophages in many parts of the body, but especially in the liver (the Kupffer cells), spleen, and bone marrow. During the next few hours to days, macrophages release the iron from the hemoglobin back into the blood to be carried by transferrin either to the bone marrow for production of new red blood cells or to the liver and other tissues for storage in the form of ferritin The porphyrin portion of the hemoglobin molecule is converted by the macrophages, through a series of stages, into the bile pigment *bilirubin*, which is released into the blood and later secreted by the liver into the bile, this will be discussed in relation to liver function in Chapter 70

THE ANEMIAS

Anemia means a deficiency of red blood cells, which can be caused either by too rapid loss or too slow production of red blood cells Some types of anemia and their physiological causes are the following

Blood Loss Anemia. After rapid hemorrhage the body replaces the plasma within one to three days, but this leaves a low concentration of red blood cells If a second hemorrhage does not occur, the red blood cell concentration returns to normal within three to four weeks

In chronic blood loss, the person frequently cannot absorb enough iron from the intestines to form hemoglobin as rapidly as it is lost Therefore, red cells are then produced with too little hemoglobin inside them, giving rise to *microcytic hypochromic anemia*, which is illustrated in Figure 4–3

Aplastic Anemia. *Bone marrow aplasia* means lack of a functioning bone marrow For instance, the person exposed to gamma ray radiation from a nuclear bomb blast is likely to sustain complete destruction of bone marrow, followed in a few weeks by lethal anemia Likewise, excessive x-ray treatment, certain industrial chemicals, and even drugs to which the person might be sensitive can cause the same effect.

Megaloblastic Anemia. From the earlier discussion in this chapter of vitamin B_{12}, folic acid, and intrinsic factor from the stomach mucosa, one can readily understand that loss of any one of these factors can lead to very slow reproduction of the erythroblasts in the bone marrow, as a result, these grow too large, with odd shapes, and are called *megaloblasts* Thus, atrophy of the stomach mucosa, as occurs in *pernicious anemia*, or loss of the entire stomach as the result of total gastrectomy can lead to megaloblastic anemia Also, patients who have intestinal sprue, in which folic acid, B_{12}, and other vitamin B compounds are poorly absorbed, often develop megaloblastic anemia Because the erythroblasts cannot proliferate rapidly enough to form normal numbers of red blood cells, the cells that are formed are mostly oversized, of bizarre shapes, and have fragile membranes Therefore, these cells rupture easily, leaving the person in dire need of an adequate number of red cells

Hemolytic Anemia. Many different abnormalities of the red blood cells, many of which are hereditarily acquired, make the cells very fragile, so that they rupture easily as they go through the capillaries, especially through the spleen Therefore, even though the number of red blood cells formed is completely normal, or even vastly excessive in some hemolytic diseases, the red cell life span is so short that serious anemia results Some of these types of anemia are the following

In *hereditary spherocytosis* the red cells are very small in size, and they are *spherical* in shape rather than being biconcave discs. These cells cannot be compressed because they do not have the normal loose, bag-like cell membrane structure of the biconcave discs Therefore, on passing through the splenic pulp they are easily ruptured by even slight compression

In *sickle cell anemia*, which is present in 0 3 to 1 0 per cent of West African and American blacks, the cells contain an abnormal type of hemoglobin called *hemoglobin S*, caused by abnormal composition of the beta chains of the hemoglobin, as explained earlier in the

chapter. When this hemoglobin is exposed to low concentrations of oxygen, it precipitates into long crystals inside the red blood cell. These crystals elongate the cell and give it the appearance of being a sickle rather than a biconcave disc. The precipitated hemoglobin also damages the cell membrane so that the cells become highly fragile, leading to serious anemia. Such patients frequently go into a vicious circle called a sickle cell disease "crisis" in which low oxygen tension in the tissues causes sickling, which causes impediment of blood flow through the tissues, in turn causing still further decrease in oxygen tension. Thus, once the process starts, it progresses rapidly, leading to a serious decrease in red blood cell mass within a few hours and, often, to death.

Thalassemia, which is also known as *Cooley's anemia* or *Mediterranean anemia*, is another hereditary type of anemia in which the cells are unable to synthesize adequate amounts of either the alpha or the beta polypeptide chains required to form the hemoglobin. Therefore, hemoglobin synthesis is greatly depressed. The hemopoietic system responds by producing tremendous numbers of new red blood cells but with little hemoglobin and also cells that are small and are fragile because much of the hemoglobin that is present is precipitated in the form of hard crystals. Here again, the cells are easily ruptured upon passing through the tissues, and many are destroyed even before they leave the bone marrow.

In *erythroblastosis fetalis*, Rh positive red blood cells in the fetus are attacked by antibodies from an Rh negative mother. These antibodies make the cells fragile and cause the child to be born with serious anemia. This will be discussed in detail in Chapter 7 in relation to the Rh factor of blood.

Hemolysis also occasionally results from transfusion reactions, from malaria, from reactions to certain drugs, and as an autoimmune process.

EFFECTS OF ANEMIA ON THE CIRCULATORY SYSTEM

The viscosity of the blood, which will be discussed in detail in Chapter 18, depends almost entirely on the concentration of red blood cells. In severe anemia the blood viscosity may fall to as low as 1.5 times that of water rather than the normal value of approximately 3 times the viscosity of water. The greatly decreased viscosity decreases the resistance to blood flow in the peripheral vessels so that far greater than normal quantities of blood return to the heart. Moreover, hypoxia due to diminished transport of oxygen by the blood causes the tissue vessels to dilate, allowing further increase in the return of blood to the heart, increasing the cardiac output to a still higher level. Thus, one of the major effects of anemia is greatly *increased work load on the heart*.

The increased cardiac output in anemia partially offsets many of the effects of anemia, for, even though each unit quantity of blood carries only small quantities of oxygen, the rate of blood flow may be increased to such an extent that almost normal quantities of oxygen are delivered to the tissues. However, when a person begins to exercise, the heart is not capable of pumping much greater quantities of blood than it is already pumping. Consequently, during exercise, which greatly increases the tissue demand for oxygen, extreme tissue hypoxia results, and acute cardiac failure often ensues.

POLYCYTHEMIA

Secondary Polycythemia. Whenever the tissues become hypoxic because of too little oxygen in the atmosphere, such as at high altitudes, or because of failure of delivery of oxygen to the tissues, as occurs in cardiac failure, the blood-forming organs automatically produce large quantities of red blood cells. This condition is called *secondary polycythemia*, and the red cell count commonly rises to as high as 6 to 8 million per cubic millimeter.

A very common type of secondary polycythemia, called *physiologic polycythemia*, occurs in natives who live at altitudes of 14,000 to 17,000 feet. The blood count is generally 6 to 8 million per cubic millimeter; this is associated with an ability of these persons to perform high levels of continuous work even in the rarefied atmosphere.

Polycythemia Vera (Erythremia). In addition to those persons who have physiologic polycythemia, others have a condition known as *polycythemia vera* in which the red blood cell count may be as high as 7 to 8 million and the hematocrit as high as 60 to 70 per cent. Polycythemia vera is a tumorous condition of the organs that produce blood cells. It causes excess production of red blood cells in the same manner that a tumor of a breast causes excess production of a specific type of breast cell. It usually also causes excess production of white blood cells and platelets.

In polycythemia vera not only does the hematocrit increase but the total blood volume also increases, rarely to almost twice normal. As a result, the entire vascular system becomes intensely engorged. In addition many of the capillaries become plugged by the viscous blood, for the viscosity of the blood in polycythemia vera sometimes increases from the normal of three times the viscosity of water to 10 times that of water.

EFFECT OF POLYCYTHEMIA ON THE CIRCULATORY SYSTEM

Because of the greatly increased viscosity of the blood in polycythemia, the flow of blood through the vessels is often very sluggish. In accordance with the factors that regulate the return of blood to the heart, as discussed in Chapter 23, it is obvious that increasing the viscosity tends to decrease the rate of venous return to the heart. On the other hand, the blood volume is greatly increased in polycythemia, which tends to increase the venous return. Actually, the cardiac output in polycythemia is not far from normal because these two factors more or less neutralize each other.

The arterial pressure is normal in most persons with polycythemia, though in approximately one third of them the pressure is elevated. This means that the blood pressure-regulating mechanisms can usually offset the tendency for increased blood viscosity to increase peripheral resistance and thereby to increase arterial pressure. Yet, beyond certain limits, these regulations fail.

The color of the skin depends to a great extent on the quantity of blood in the subpapillary venous [illegible]

polycythemia vera the quantity of blood in this plexus is greatly increased. Furthermore, because the blood passes sluggishly through the skin capillaries before entering the venous plexus, a larger than normal proportion of the hemoglobin is deoxygenated before the blood enters the plexus The blue color of this deoxygenated hemoglobin masks the red color of the oxygenated hemoglobin. Therefore, a person with polycythemia vera ordinarily has a ruddy complexion but often with a bluish (cyanotic) tint to the skin. (In secondary polycythemia cyanosis is almost always evident because hypoxia is the usual cause of this type of polycythemia)

REFERENCES

Baum S J and Ledney, G D (eds) Experimental Hematology Today 1978 New York, Springer-Verlag, 1978

Benz, E H Jr , and Forget, B G The thalassemia syndromes Models for the molecular analysis of human disease *Annu Rev Med* 33 363, 1982

Bessis, M , *et al* (eds) Red Cell Rheology New York Springer Verlag 1978

Beutler, E Hemolytic Anemia in Disorders of Red Cell Metabolism New York, Plenum Press, 1978

Bottomley S S , and Whitcomb W H Erythropoiesis *In* Fröhlich E D (ed) Pathophysiology, 2nd Ed Philadelphia J B Lippincott Co 1976, p 567

Brain, M C Current Therapy in Hematology-Oncology 1983 1984 St Louis, The C V Mosby Co , 1983

Charache, S Treatment of sickle cell anemia *Annu Rev Med* , 32 195, 1981

Charlton R W , and Bothwell T H Iron absorption *Annu Rev Med* 34 55, 1983

Cokelet, G R *et al* (eds) Erythrocyte Mechanics and Blood Flow New York, A R Liss 1979

Dunn, C D R The Differentiation of Haemopoietic Stem Cells Baltimore, Williams & Wilkins, 1971

Eaves A C and Eaves, C J Erythropoiesis in culture *Clin Haematol* , 13 371, 1984

Eaves, C J , and Eaves A C Factors and hemopoiesis Pandora s box revisited *Prog Clin Biol Res* , *148* 83, 1984

Ellory C and Lew V L (eds) Membrane Transport in Red Cells New York, Academic Press 1977

Erslev, A J , and Gabuzda T G Pathophysiology of Blood Philadelphia W B Saunders Co , 1979

Fevery J and Heirwegh, K P M Bilirubin metabolism *In* Javitt N B (ed) International Review of Physiology Liver and Biliary Tract Physiology I Vol 21 Baltimore University Park Press, 1980 p 171

Freedman, M H Congenital failure of hematopoiesis in the newborn infant *Clin Perinatol* 11 417 1984

Fung Y C Red blood cells and their deformability *In* Kaley G and Altura B M (eds) Microcirculation Vol I Baltimore, University Park Press, 1977 p 255

Hansen, R M , *et al*., Failure to suspect and diagnose thalassemic syndromes Interpretation of RBC indices by the nonhematologist *Arch Intern Med* 145 93, 1985

Harrison, P R Molecular analysis of erythropoiesis A current appraisal *Exp Cell Res* 155 321, 1984

Heimpel H *et al* (eds) International Symposium on Aplastic Anemia New York Springer Verlag, 1979

Hoffbrand A V , *et al* (eds) Recent Advances in Haematology New York Longman Inc 1977

Jacquez J A Red blood cell as glucose carrier Significance for placenta and cerebral glucose transfer *Am J Physiol* , 246 R289, 1984

Johannessen, J V Electron microscopy in human medicine *In* Cardiovascular System Lymphoreticular and Hematopoietic System V New York, McGraw Hill Book Co , 1982

Johnson, G R Haemopoietic multipotential stem cells in culture *Haematol* , 13 309, 1984

Kass, L Bone Marrow Interpretation. Philadelphia J B Lippincott Co , 1979

Kelemen, E , *et al* Atlas of Human Hemopoietic Development. New York, Springer Verlag 1979

Konigsberg W Protein structure and molecular dysfunction Hemoglobin *In* Bondy, P K , and Rosenberg, L E (eds) Metabolic Control and Disease, 8th Ed Philadelphia W B Saunders Co , 1980, p

Kruckeberg W C , *et al* (eds) Erythrocyte Membranes New York R Liss 1978

Lux, S E et *al* (eds.) Normal and Abnormal Red Cell Membranes New York, A R Liss, 1979

Macey, R I Transport of water and urea in red blood cells. *Am J Physiol* , *246* C195, 1984

Mackinney A A Jr The Pathophysiology of Blood New York, John Wiley & Sons, 1984

Maclean, N Haemoglobin London, Edward Arnold, 1978

Maslow, W C , *et al* Hematologic Disease Boston Houghton Mifflin 1979

Metcalf D and Nicola N A The regulatory factors controlling mouse erythropoiesis in vitro *Prog Clin Biol Res* *148* 93 1984

Munro H N , and Linder, M C Ferritin Structure biosynthesis, and role in iron metabolism *Physiol Rev* 58 317, 1978

Ogawa M *et al* Differentiation and proliferative kinetics of hematopoietic stem cells in culture *Prog Clin Biol Res* , *148* 35 1984

Peschle, C Erythropoiesis *Annu Rev Med* 31 303 1980

Petz, L D Red cell transfusion problems in immunohematologic disease *Annu Rev Med* , *33* 355 1982

Petz, L D , and Garratty, G Acquired Immune Hemolytic Anemia New York Churchill Livingstone, 1979

Platt W R Color Atlas and Textbook of Hematology Philadelphia J B Lippincott Co , 1978

Propper, R and Nathan, D Clinical removal of iron *Annu Rev Med* 33 509 1982

Schmid Schonbein H *et al* Biology of red cells Non nucleated erythrocytes as fluid drop-like cell fragments *Int J Microcirc Clin Exp* 3 161, 1984

Seiverd, C E Hematology for Medical Technologists Philadelphia, Lea & Febiger, 1983

Spivak J L (ed) Fundamentals of Clinical Hematology Hagerstown Md , Harper & Row 1980

Stamatoyannopoulos G , and Nienhuis A W (eds) Cellular and Molecular Regulation of Hemoglobin Switching New York, Grune & Stratton 1979

Stanley, E R , and Jubinsky, P T Factors affecting the growth and differentiation of haemopoietic cells in culture *Clin Haematol* 13 329 1984

Swisher, S N (ed) Immune Hemolytic Anemias New York, Grune & Stratton, 1976

Thompson, R B A Short Textbook of Haematology Baltimore, Williams & Schwarzenberg, 1984

Torok Storb B T-cell effects on in vitro erythropoiesis Immune regulation and immune reactivity *Prog Clin Biol Res* , *148* 163, 198

Tschudy, D P , and Lamon J M Porphyrin metabolism and the porphyrias *In* Bondy, P K , and Rosenberg L E (eds) Metabolic Control and Disease, 8th Ed Philadelphia, W. B Saunders Co , 1 p 939

Weatherall, D J (ed) Advances in Red Blood Cell Biology New York Raven Press, 1982

Zagalak, B , and Friedrich, W (eds) Vitamin B12 New York, Walter De Gruyter, 1979

5

Resistance of the Body to Infection —The Leukocytes, the Macrophage System, and Inflammation

Our bodies normally are exposed to bacteria, viruses, fungi, and parasites, which occur especially in the skin, the mouth, the respiratory passageways, the intestinal tract, the lining membranes of the eyes, and even the urinary tract. Many of these agents are capable of causing serious diseasa if they invade the deeper tissues. In addition, wa ara exposed intermittently to other highly infectious bacteria and viruses besides those that are normally present in our bodies, and these cause lethal diseases such as pneumonia, streptococcal infection, and typhoid fever.

Fortunately, our bodies have a special system for combatting the different infectious and toxic agents. This is composed of the *leukocytes* (also called white blood cells), the *macrophage system* (frequently but incorrectly called the reticuloendothelial system), and the *lymphoid tissue*. These tissues function in two different ways to prevent disease: (1) by actually destroying invading agents by the process of phagocytosis and (2) by forming antibodies and sensitized lymphocytes, one or both of which may destroy the invader. The present chapter is concerned with the first of these methods, while the following chapter is concerned with the second.

THE LEUKOCYTES (WHITE BLOOD CELLS)

The leukocytes are the *mobile units* of the body's protective system. They are formed partially in the bone marrow (the *granulocytes* and *monocytes*, and a few *lymphocytes*) and partially in the lymph tissue (*lymphocytes* and *plasma cells*), but after formation they are transported in the blood to the different parts of the body where they are to be used. The real value of the white blood cells is that most of them are specifically transported to areas of serious inflammation, thereby providing a rapid and potent defense against any infectious agent that might be present. As we shall see later, the granulocytes and monocytes have a special capability to "seek out and destroy" any foreign invader

GENERAL CHARACTERISTICS OF LEUKOCYTES

The Types of White Blood Cells. Six different types of white blood cells are normally found in the blood These are *polymorphonuclear neutrophils, polymorphonuclear eosinophils, polymorphonuclear basophils, monocytes, lymphocytes* and occasional *plasma cells*. In addition, there are large numbers of *platelets*, which are fragments of a seventh type of white cell found in the bone marrow, the *megakaryocyte* The three types of polymorphonuclear cells have a granular appearance, as illustrated in Figure 5–1, for which reason they are called *granulocytes*, or in clinical terminology they are often called simply "polys"

Tha granulocytes and the monocytes protect the body against invading organisms by ingesting them—that is, by the process of *phagocytosis* The lymphocytes and plasma cells function mainly in connection with the immune system; this will be discussed in the following chapter. However, a function of certain lymphocytes is to attach to specific invading organisms and to destroy them, similar to the actions of the granulocytes and monocytes. Finally, the function of platelets is to activate the blood-clotting mechanism, which will be discussed in Chapter 8.

Concentrations of the Different White Blood Cells in the Blood. The adult human being has approximately 7000 white blood cells per cubic millimeter of blood. The normal percentages of the different types of white blood cells are approximately the following:

Polymorphonuclear neutrophils	62.0%
Polymorphonuclear eosinophils	2.3%
Polymorphonuclear basophils	0.4%
Monocytes	5.3%
Lymphocytes	30.0%

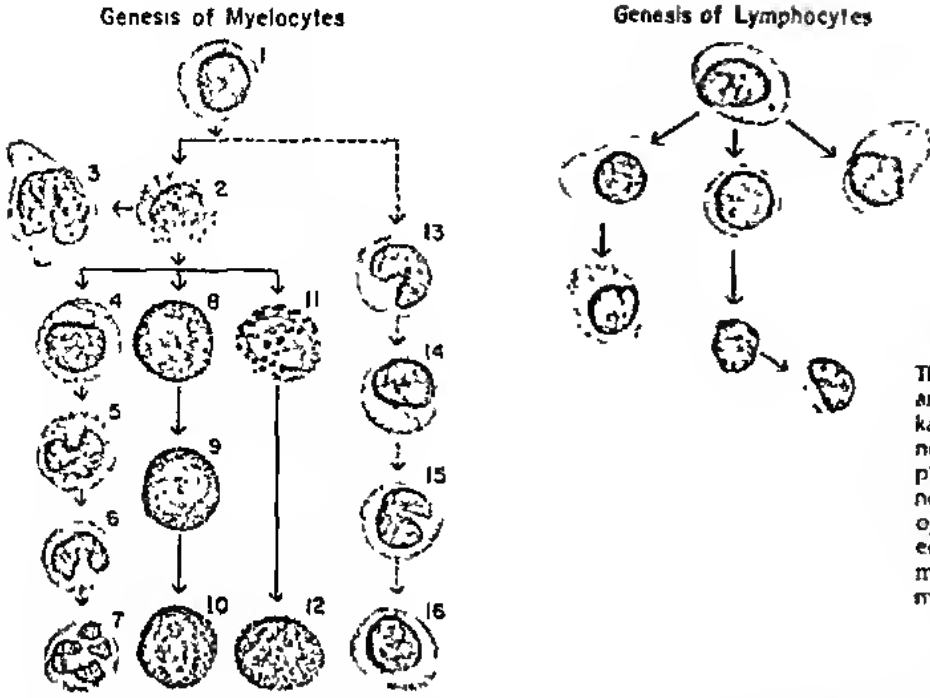


Figure 5–1. Genesis of the white blood cells. The different cells of the myelogenous series are 1, myeloblast; 2, promyelocyte; 3, megakaryocyte; 4, neutrophil myelocyte; 5, young neutrophil metamyelocyte; 6, 'band' neutrophil metamyelocyte; 7, polymorphonuclear neutrophil; 8, eosinophil myelocyte; 9, eosinophil metamyelocyte; 10, polymorphonuclear eosinophil; 11, basophil myelocyte; 12, polymorphonuclear basophil; 13–16, stages of monocyte formation.

The number of platelets, which are only cell fragments, in each cubic millimeter of blood is normally about 300,000.

GENESIS OF THE LEUKOCYTES

Figure 5–1 illustrates the stages in the development of white blood cells. The polymorphonuclear cells and monocytes are normally formed only in the bone marrow. On the other hand, lymphocytes and plasma cells are produced in the various lymphogenous organs, including the lymph glands, the spleen, the thymus, the tonsils, and various lymphoid rests in the bone marrow, gut, and elsewhere.

The white blood cells formed in the bone marrow, especially the granulocytes, are stored within the marrow until they are needed in the circulatory system. Then when the need arises, various factors that are discussed later cause them to be released. Normally, about three times as many granulocytes are stored in the marrow as circulate in the entire blood. This represents about a six day supply of granulocytes.

As illustrated in Figure 5–1, megakaryocytes are also formed in the bone marrow and are part of the myelogenous group of bone marrow cells. These megakaryocytes fragment in the bone marrow, the small fragments known as *platelets* or *thrombocytes* passing then into the blood.

Materials Needed for Formation of White Blood Cells. In general, the white blood cells need essentially the same vitamins and amino acids as most of the other cells of the body for their formation. Especially does lack of folic acid, a compound of the vitamin B complex, block the formation of white blood cells as well as prevent maturation of red blood cells, which was pointed out in Chapter 4. Also, in extreme metabolic debilitation, production of white blood cells may be greatly reduced, despite the fact that these cells are needed more than usual during such a state.

LIFE SPAN OF THE WHITE BLOOD CELLS

The main reason white blood cells are present in the blood is simply to be transported from the bone marrow or lymphoid tissue to the areas of the body where they are needed. Therefore, it is to be expected that the transit time of the white blood cells in the blood would be short.

The subsequent life of the granulocytes once released from the bone marrow is normally four to eight hours circulating in the blood and another four to five days in the tissues. In times of serious tissue infection, this total life span is often shortened to only a few hours because the granulocytes then proceed rapidly to the infected area, ingest the invading organisms, and in the process are themselves destroyed.

The monocytes also have a short transit time in the blood before wandering through the capillary membranes into the tissues. However, once in the tissues they swell to much larger sizes to become tissue macrophages and in this form can live for months or even years unless destroyed by performing phagocytic function. These tissue macrophages form the basis of the tissue macrophage system that provides a first line of defense in the tissues against infection, as we shall discuss later in the chapter.

Lymphocytes enter the circulatory system continually along with the drainage of lymph from the lymph nodes. The total number entering the blood from the thoracic duct in each 24 hours is usually several times the total number of lymphocytes present in the blood stream at any given time. Therefore, the span of time that the lymphocytes remain in the blood must be only a few hours. However, studies using radioactive lymphocytes have shown that almost all of these pass by diapedesis back into the tissues, then re-enter the lymph and return to the blood again and again, thus, there is continual circulation of the lymphocytes through the tissues. And many of these cells have life spans of 100 to 300 days.

or in some instances perhaps even years, but this also depends on the body's need for these cells.

The platelets in the blood are totally replaced approximately once every 10 days; in other words, about 30,000 platelets are formed each day for each cubic millimeter of blood.

PROPERTIES OF NEUTROPHILS, MONOCYTES, AND MACROPHAGES

It is mainly the neutrophils and the monocytes that attack and destroy invading bacteria, viruses, and other injurious agents. The neutrophils are mature cells that can attack and destroy bacteria and viruses even in the circulating blood. On the other hand, the blood monocytes are immature cells that have very little ability to fight infectious agents. However, once they enter the tissues they begin to swell, often increasing their diameters as much as fivefold, to as great as 80 microns, a size that can be seen with the naked eye. Also, extremely large numbers of lysosomes and mitochondria develop in the cytoplasm, giving the cytoplasm the appearance of a bag filled with granules. These cells are now called *macrophages*, and they are extremely capable of combating disease agents.

Diapedesis. Neutrophils and monocytes can squeeze through the pores of the blood vessels by the process of diapedesis. That is, even though a pore is much smaller than the size of the cell, a small portion of the cell slides through the pore at a time, the portion sliding through being momentarily constricted to the size of the pore, as illustrated in Figure 5–2.

Amebold Motion. Both neutrophils and macrophages move through the tissues by ameboid motion, which was described in Chapter 2. Some of the cells can move through the tissues at rates as great as 40 microns per minute—that is, they can move several times their own length each minute.

Chemataxia. A number of different chemical substances in the tissues cause both neutrophils and macrophages to move either toward or away from the source of the chemical. This phenomenon is known as *chemotaxis*. When a tissue becomes inflamed, a number of different products can cause chemotaxis of both neutrophils and macrophages, causing them to move toward the inflamed area. These include (a) some of the bacterial toxins, (b) degenerative products of the inflamed tissues themselves, (c) several reaction products of the "complement complex" (discussed in the following chapter), (d) several reaction products caused by plasma clotting in the inflamed area, and (e) still other substances.

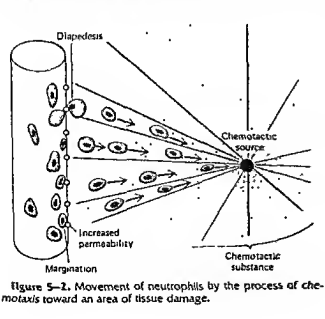


Figure 5–2. Movement of neutrophils by the process of *chemotaxis* toward an area of tissue damage.

As illustrated in Figure 5–2, chematoxis depends on a concentration gradient of the chemotactic substance. The concentration is greatest near the source, which causes directional movement of the leukocytes—including the neutrophils, monocytes, and other white blood cells as well—toward the inflamed area. Chematoxis is very effective up to 100 microns away from an inflamed tissue, since almost no tissue area is more than 30 to 50 microns away from a capillary, the chemotactic signal can easily move vast hordes of leukocytes from the capillaries into the inflamed area.

Margination and Diapedesis of Leukocytes Through the Capillary Membrane. Some of these same products that cause chemotaxis, as well as enzymes and necrotic products released from the inflamed tissues, have a direct effect on the local capillaries to cause margination and diapedesis of white blood cells. Margination means sticking of the white blood cells to the wall. Even normally, granulocytes and monocytes, partly because of their large sizes and partly because of a natural stickiness, adhere to the capillary walls. In fact, about three fifths of the granulocytes and three fourths of the monocytes in the blood are sequestered in the capillaries in this manner. However, after the inflammatory products have acted on the capillary walls, the endothelial cells become especially sticky to the white blood cells, causing extreme amounts of margination of both the granulocytes and monocytes, as illustrated in Figure 5–2. At the same time, the pores of the capillaries open more widely than normally to allow rapid diapedesis of the cells into the tissues, and thenca chemotaxis toward tha inflamed site.

Phagocytasis. The most important function of the neutrophils and macrophages is phagocytosis.

Obviously, tha phagocytes must be selective of tho material that is phagocytized, or otherwise some of the normal cells and structures of tbe body would be ingested. Whether or not phagocytosis will occur depends especially upon three selective procedures. First, if the surface of a particle is rough, the likelihood of phagocytosis is increased. Second, most natural substances of the body have protective protein coats that repel the phagocytes. On the other hand, dead tissues and foreign particles frequently have no protective coats, and many are also strongly electrically charged, which also makes them subject to phagocytosis. Third, the body has a specific means for recognizing certain foreign materials. This is the function of the immune system that will be described in the following chapter. The immune system develops antibodies against infectious agents like bacteria. These antibodies then adhere to the bacterial membranes and thereby make the bacteria especially susceptible to phagocytosis This results from the antibody molecule also binding with another factor called *complement*, which is an additional part of the immune system as will be discussed in the following chapter. Some of the products of complement then "cover" the bacteria and make them very susceptible to phagocytosis, which is the process known as *opsonization*. Complement can also attach to some bacteria even in the

absence of antibodies, and this too can lead to opsonization.

Phagocytosis by Neutrophils. The neutrophils entering the tissues are already mature cells that can immediately begin phagocytosis On approaching a particle to be phagocytized, the neutrophil first attaches to receptors on the particle, then projects pseudopodia in all directions around the particle, and the pseudopodia meet each other on the opposite side and fuse. This creates an enclosed chamber containing the phagocytized particle. Then the chamber invaginates to the inside of the cytoplasmic cavity and breaks away from the outer cell membrane to form a free-floating *phagocytic vesicle* (also called a *phagosome*) inside the cytoplasm

A neutrophil can usually phagocytize 5 to 20 bacteria before the neutrophil itself becomes inactivated and dies

Phagocytosis by Macrophages. Macrophages, when activated by the immune system as described in the following chapter, are much more powerful phagocytes than the neutrophils, often capable of phagocytizing as many as 100 bacteria They have the ability to engulf much larger particles and often five or more times as many particles as the neutrophils And they can even phagocytize whole red blood cells or malarial parasites, whereas neutrophils are not capable of phagocytizing particles much larger than bacteria Also, macrophages have the ability to phagocytize necrotic tissue and even dead neutrophils, which is a very important function performed by these cells in chronic infection

Enzymatic Digestion of the Phagocytized Particles. Once a foreign particle has been phagocytized, lysosomes immediately come in contact with the phagocytic vesicle, and their membranes fuse with those of the vesicle, thereby dumping many digestive enzymes of the lysosomes into the vesicle Thus, the phagocytic vesicle now becomes a *digestive vesicle*, and digestion of the phagocytized particle begins immediately

Neutrophils and macrophages both have an abundance of lysosomes filled with *proteolytic enzymes* especially geared for digesting bacteria and other foreign protein matter The lysosomes of macrophages also contain large amounts of *lipases*, which digest the thick lipid membranes possessed by tubercle bacteria, leprosy bacteria, and others.

The Ability of Neutrophils and Macrophages to Kill Bacteria. In addition to lysosomal digestion of ingested bacteria in the phagosomes, neutrophils and macrophages also contain bactericidal agents that kill most bacteria even when the lysosomal enzymes fail to digest them This is especially important because some bacteria have protective coats or other factors that prevent their destruction by the enzymes Most of the killing effect results from several powerful *oxidizing agents* formed by enzymes in the membrane of the phagosome or by the special organelle called the *peroxisome* These oxidizing agents include large quantities of *superoxide* (O_2), *hydrogen peroxide* (H_2O_2), and *hydroxyl ions* (—OH), all of which are lethal to most bacteria even in very small quantities

Another lysosomal substance that is lethal to many bacteria in the phagosome is *lysozyme*, a chemical that can cause dissolution of the lipid membranes of the bacteria

Unfortunately, though, some bacteria, notably the tubercle bacillus, have coats that are resistant to lysosomal digestion and at the same time secrete substances that resist even the killing effects of the neutrophils and macrophages. Often these bacteria are responsible for many of the chronic diseases.

Special Functions of Macrophages in Relation to the Immune System. In the following chapter we shall discuss some of the special relationships that macrophages have with the immune system. The two most important of these are:

(1) Macrophages in the lymphoid tissue make direct contact with many of the lymphocytes. When the macrophages ingest foreign organisms, they pass the antigenic material derived from these organisms to the lymphocytes and activate the lymphocytes to begin producing antibodies.

(2) Another product of the immune system is sensitized lymphocytes, called "T cells," that can destroy foreign organisms and can also initiate inflammatory reactions in infected tissues. Some of the products of these T cells cause macrophages to migrate into the inflamed tissues and also strongly activate the macrophages, thus helping combat the infection These functions will be discussed further in the following chapter.

Death of the Neutrophils and Macrophages as a Result of Phagocytosis. Neutrophils continue to ingest and digest foreign particles until toxic substances from these, as well as enzymes released inside the neutrophils from the lysosomes, kill the neutrophils themselves This usually occurs after a neutrophil has phagocytized from 5 to 25 bacteria. Then the macrophages set about phagocytizing and digesting the dead neutrophils

The macrophages often also are killed in their process of phagocytizing infectious agents. However, this is not always the case because the macrophage, in contradistinction to the neutrophil, is capable of extruding the residual breakdown products of the phagocytized agent and therefore continuing to perform its function sometimes for weeks, months, or perhaps even years.

THE TISSUE MACROPHAGE SYSTEM (THE RETICULOENDOTHELIAL SYSTEM)

In the above paragraphs we have described the macrophages mainly as mobile cells that are capable of wandering through the tissues. However, a large portion of the monocytes, on entering the tissues and after becoming macrophages, become attached to the tissues and remain attached for months or perhaps even years unless they are called upon to perform specific protective functions. They have the same capabilities as the mobile macrophages to phagocytize large quantities of bacteria, viruses, necrotic tissue, or other foreign particles in the tissue And, when appropriately stimulated, they can break away from their attachments and become mobile macrophages that respond to chemotaxis and all the other stimuli related to the inflammatory process.

The combination of mobile macrophages and fixed tissue macrophages is collectively called the *reticuloendothelial system.* The reason for this name is that it was formerly believed that a major share of the blood vessel endothelial cells could perform phagocytic functions similar to those performed by the macrophage system However, recent studies have disproved this. Therefore, the reticuloendothelial system is actually a misnomer. Yet, because the term is so widely used, it should be remembered to be almost synonymous with the tissue macrophage system

The tissue macrophages in various tissues differ in appearances because of environmental differences, and

they are known by different names: *Kupffer cells* in the liver; *tissue macrophages* in the lymph nodes, spleen, and bone marrow; *alveolar macrophages* in the alveoli of the lungs; *tissue histiocytes, clasmatocytes,* or *fixed macrophages* in the subcutaneous tissues; and *microglia* in the brain. Let us describe briefly the function of the tissue macrophages in some areas of the body that are especially exposed to infectious agents.

Tissue Macrophages in the Skin and Subcutaneous Tissues (Histiocytes). Though the skin is normally impregnable to infectious agents, this no longer holds true when the skin is broken. When infection does begin in the subcutaneous tissues and local inflammation ensues, the tissue macrophages can divide *in situ* and form more macrophages. Then they perform the usual functions of attacking and destroying the infectious agents, as described earlier.

Macrophages of the Lymph Nodes. Essentially no particulate matter that enters the tissues can be absorbed directly through the capillary-membranes into the blood. Instead, if the particles are not destroyed locally in the tissues, they enter the lymph and flow through the lymphatic vessels to the lymph nodes located intermittently along the course of the lymphatic. The foreign particles are trapped there in a meshwork of sinuses lined by *tissue macrophages.*

Figure 5–3 illustrates the general organization of the lymph node, showing lymph entering by way of the *afferent lymphatics,* flowing through the *medullary sinuses,* and finally passing out of the *hilus* into the *efferent lymphatics.* Large numbers of macrophages line the sinuses, and if any particles enter the sinuses the macrophages phagocytize them and prevent general dissemination throughout the body.

Alveolar Macrophages. Another route by which invading organisms frequently enter the body is through the respiratory system. Fortunately, large numbers of tissue macrophages are present as integral components of the alveolar walls. These can phagocytize particles that become entrapped in the alveoli. If the particles are digestible, the macrophages can also digest them and release the digestive products into the lymph. If the particle is not digestible the macrophages often form a "giant cell" capsule around the particle until such time, if ever, that it can be slowly dissoluted. Such capsules are frequently formed around tubercle bacilli, silica dust particles, and even carbon particles.

Tissue Macrophages (Kupffer Cells) in the Liver Sinuses. Still another favorite route by which bacteria

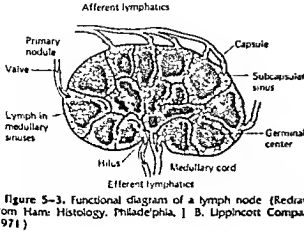


Figure 5–3. Functional diagram of a lymph node (Redrawn from Hamerson Histology. Philadelphia, J. B. Lippincott Company, 1971.)

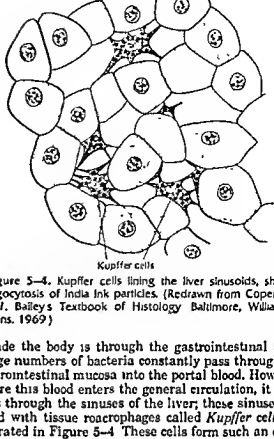


Figure 5–4. Kupffer cells lining the liver sinusoids, showing phagocytosis of India ink particles. (Redrawn from Copenhaver et al. Bailey's Textbook of Histology. Baltimore, Williams & Wilkins, 1969.)

invade the body is through the gastrointestinal tract. Large numbers of bacteria constantly pass through the gastrointestinal mucosa into the portal blood. However, before this blood enters the general circulation, it must pass through the sinuses of the liver; these sinuses are lined with tissue macrophages called *Kupffer cells,* illustrated in Figure 5–4. These cells form such an effective particulate filtration system that almost none of the bacteria from the gastrointestinal tract succeeds in passing from the portal blood into the general systemic circulation. Indeed, motion pictures of phagocytosis by Kupffer cells have demonstrated phagocytosis of a single bacterium in less than 1/100 second.

Macrophages of the Spleen and Bone Marrow. If an invading organism does succeed in entering the general circulation, there still remain other lines of defense by the tissue macrophage system, especially by macrophages of the spleen and bone marrow. In both these tissues, macrophages have become entrapped by the reticular meshwork of the two organs, and when foreign particles come in contact with these they are phagocytized.

The spleen is similar to the lymph nodes, except that blood, instead of lymph, flows through the substance of the spleen. Figure 5–5 illustrates the spleen's general structure, showing a small peripheral segment. Note that a small artery penetrates from the splenic capsule into the *splenic pulp,* and terminates in small capillaries. The capillaries are highly porous, allowing large numbers of whole blood cells to pass out of the capillaries into the *cords of the red pulp.* These cells then gradually *squeeze* through the trabecular meshwork of the cords and eventually return to the circulation through the endothelial walls of the venous sinuses. The trabeculae of the red pulp are lined with vast numbers of macrophages, and in addition the venous sinuses are also lined with macrophages. This peculiar passage of blood through the cords of the red pulp provides an exceptional means for phagocytosis of unwanted debris in the blood, especially old and abnormal red blood cells. The spleen is also an important organ for phagocytic removal of abnormal platelets, blood parasites, and any bacteria

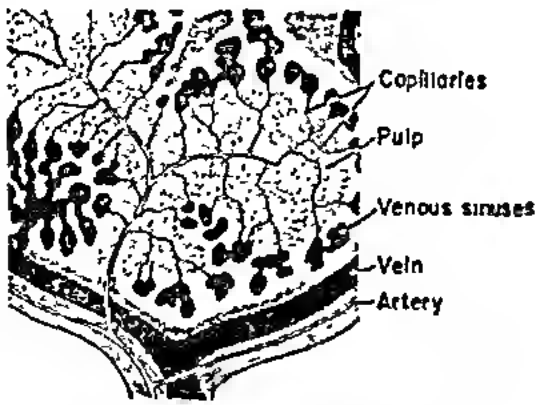


Figure 5–5. Functional structures of the spleen. (Modified from Bloom and Fawcett Textbook of Histology Philadelphia, W B Saunders Company 1975)

that might succeed in entering the general circulating blood

In a similar way, macrophages of the bone marrow also help remove unwanted debris and pathologic agents from the blood

INFLAMMATION AND FUNCTION OF NEUTROPHILS AND MACROPHAGES

THE PROCESS OF INFLAMMATION

When tissue injury occurs, whether it be caused by bacteria, trauma, chemicals, heat, or any other phenomenon, multiple substances that cause dramatic secondary changes in the tissues are released by the injured tissues These secondary changes are called *inflammation* Inflammation is characterized by (1) vasodilatation of the local blood vessels with consequent excess local blood flow, (2) increased permeability of the capillaries with leakage of large quantities of fluid into the interstitial spaces, (3) often clotting of the fluid in these spaces because of excessive amounts of fibrinogen and other proteins leaking from the capillaries, and (4) swelling of the cells Some of the many tissue products that cause these reactions include *histamine, bradykinin, serotonin, prostaglandins*, several different *reaction products of the complement system* (that will be described in the following chapter), *reaction products of the blood-clotting system*, and multiple *hormonal substances called* "*lymphokines*" that are released by sensitized T cells, part of the immune system, as we shall also discuss in the following chapter Several of these substances strongly activate the machrophage system, and within a few hours the macrophages begin to devour the destroyed tissues—but at times the macrophages also further injure the still-living tissue cells

The "Walling Off" Effect of Inflammation. One of the first results of inflammation is to "wall off" the area of injury from the remaining tissues The tissue spaces and the lymphatics in the inflamed area are blocked by fibrinogen clots so that fluid barely flows through the spaces Therefore, this walling-off process delays the spread of bacteria or toxic products

The intensity of the inflammatory process is usually proportional to the degree of tissue injury For instance, staphylococci invading the tissues liberate extremely lethal cellular toxins. As a result, the process of inflammation develops rapidly—indeed, much more rapidly than the staphylococci themselves can multiply and spread. Therefore, staphylococcal infection is characteristically walled off rapidly. On the other hand, streptococci do not cause such intense local tissue destruction. Therefore the walling off proress develops slowly while the streptococci reprodure and migrate As a result, streptococci have a far greater tendency to spread through the body and cause death than do staphylococci, even though staphylococci are far more destructive to the tissues

The Macrophage and Neutrophil Response to Inflammation

Soon after inflammation begins, the inflamed area becomes invaded by both neutrophils and macrophages, and these set about performing their scavenger functions to rid the tissue of infectious or toxic agents However, the macrophage and neutrophil responses occur in several different stages

The Tissue Macrophages as the First Line of Defense. The macrophages that are already present in the tissues, whether they be the histiocytes in the subcutaneous tissues, the alveolar macrophages in the lungs, the microglia in the brain, or others, immediately begin their phagocytic actions Therefore, they are the first line of defense against infection during the first hour or so However, their numbers often are not very great

Neutrophilia and Neutrophil Invasion of the Inflamed Area—The Second Line of Defense. The term *neutrophilia* means an increase in the number of neutrophils in the blood. The term *leukocytosis* is also often used to mean the same as neutrophilia, though this term actually means excess number of all white cells, whatever their types

Within a few hours after the onset of acute inflammation, the number of neutrophils in the blood sometimes increases as much as four- to fivefold—to as high as 15,000 to 25,000 per cubic millimeter This results from a combination of chemical substances that are released from the inflamed tissues, collectively called *leukocytosis-inducing factor* or *granulocyte-releasing factor* This factor diffuses from the inflamed tissue into the blood and is carried to the bone marrow There it mobilizes large numbers of leukocytes, mainly neutrophils, that are already preformed and stored in the marrow tissue In this way large numbers of neutrophils are almost immediately transferred from the bone marrow storage pool into the circulating blood

Movement of Neutrophils to the Area of Inflammation. Products from the inflamed tissues also cause neutrophils to move from the circulation into the inflamed area They do this in three ways

First, they damage the capillary walls and thereby cause neutrophils to stick, which is the process called *margination* that was illustroted in Figure 5–2

Second, they greatly increase the permeability of the capillaries and small venules, and this allows the neutrophils to pass by *diapedesis* into the tissue spaces

Third, there is also the phenomenon of *chemotaxis* that causes the neutrophils to migrate toward the injured tissues This is caused by bacterial or cellular products, clotting reaction factors, antigen-antibody products, and some of the "complement" components that will be described in the following chapter.

Thus, within several hours after tissue damage begins, the area becomes well supplied with neutrophils. Since neutrophils are already mature cells, they are ready to begin immediately their scavenger functions for removal of foreign matter from the inflamed tissues.

Increased Production of Neutrophils by the Bone Marrow—Colony-Stimulating Factor. Inflammation not only causes release of leukocytes from the bone marrow's storage pool but causes an increased rate of production of leukocytes as well. A number of different factors seem to cause this; several of these have been isolated and characterized and collectively are called *colony-stimulating factor* because they all cause colonies of leukocytes to proliferate in tissue culture as well as causing increased leukocyte production in the bone marrow. Colony-stimulating factor is a collection of glycoproteins with molecular weights of between 20,000 and 50,000. One of the most abundant of them is produced by the macrophages in the inflamed tissues, probably in response to bacterial toxins, breakdown products of neutrophils, and other products of the inflamed tissues

Once released from the macrophages, colony-stimulating factor acts mainly on the earliest ancestors of the white blood cells, causing especially rapid reproduction of the granulopoietic stem cells and myeloblasts. Thus, stimulation of increased production of granulocytes is important for maintaining a high level of neutrophilia for many days, weeks, or even months—but not for the early stages of neutrophilia.

Macrophage Proliferation and the Monocyte Response—The Third Line of Defense. Though the first and second lines of defense in tissue infection are the already present tissue macrophages and the rapid production and movement of neutrophils into the inflamed area, still a third line of defense is a slow but long-continuing increase in the number of macrophages as well. This results partly from reproduction of the already present tissue macrophages but also from migration of large numbers of monocytes into the inflamed area. Though the monocytes are still immature cells and do not have the capability of phagocytosis when they first enter the tissues, over a period of 8 to 12 hours they swell markedly, form greatly increased quantities of cytoplasmic lysosomes, exhibit increased ameboid motion, and move chemotactically toward the damaged tissues.

Next, the rate of production of monocytes by the bone marrow also increases. This possibly results partly from stimulation by colony-stimulating factor, but there seem to be other, as yet undefined stimulating factors as well, because in long-term chronic infection there is progressively increasing production of monocytes, thereby increasing the ratio of macrophages to neutrophils in the tissues. Therefore, the long-term chronic defense against infection is often mainly a macrophage response rather than a neutrophil response.

As has already been pointed out, the macrophages can phagocytize far more bacteria and far larger particles, including even neutrophils and large quantities of necrotic tissue. Also, the macrophages play an important role in initiating the development of antibodies, as we shall discuss in the following chapter.

Formation of Pus

When the neutrophils and macrophages engulf large numbers of bacteria and necrotic tissue, essentially all the neutrophils and many if not most of the macrophages themselves eventually die. After several days, a cavity is often excavated in the inflamed tissues containing varying portions of necrotic tissue, dead neutrophils, and dead macrophages Such a mixture is commonly known as *pus*.

Ordinarily, pus formation continues until all infection is suppressed. Sometimes the pus cavity eats its way to the surface of the body or into an internal cavity and in this way empties itself. At other times the pus cavity remains closed even after tissue destruction has ceased. When this happens the dead cells and necrotic tissue in the pus gradually autolyze over a period of days, and the end-products of autolysis are usually absorbed into the surrounding tissues until most of the evidence of tissue damage is gone.

NEUTROPHILIA CAUSED BY CONDITIONS OTHER THAN INFLAMMATION

Almost any factor that causes some degree of tissue destruction will cause neutrophilia For instance, persons debilitated by cancer exhibit an increase in neutrophils from the normal of 4500 per cubic millimeter sometimes up to 15,000 or more. Even extreme fatigue can cause neutrophilia Acute hemorrhage, poisoning, operative procedures, very slight hemorrhage into the peritoneal cavity, and injection of foreign protein into the body all cause considerable increase in the number of neutrophils in the circulatory system. In summary, neutrophilia results from almost any tissue-damaging process in the body, whether or not this process is associated with inflammation

For instance, neutrophilia is one of the diagnostic features of coronary thrombosis. Presumably, when the coronary vessel becomes blocked the ischemic musculature of the heart begins to necrose, and degenerative substances liberated into the blood promote the release of neutrophils from the bone marrow

Physiological Neutrophilia. The number of neutrophils in the circulatory system can increase to as much as two to three times normal after a single minute of extremely hard exercise or after injection of norepinephrine. This can be explained as follows: When blood flow is sluggish through the tissues, large numbers of white blood cells, especially neutrophils, adhere to the walls of the capillaries—a process called *margination*—and, therefore, are sequestered from the usual circulation. Hard exercise or stimulation of the circulation by norepinephrine, with rapid flow of blood through essentially all capillaries, can mobilize the leukocytes.

Approximately one hour after physiological neutrophilia has resulted from exercise or any other stimulus, the number of leukocytes in the blood is usually back to normal because most of the leukocytes will again be sequestered in the capillaries.

THE EOSINOPHILS

The eosinophils normally comprise 2 to 3 per cent of all the blood leukocytes Eosinophils are weak phagocytes, and they exhibit chemotaxis, but in comparison with the neutrophils it is doubtful that the eosinophils are of significant importance in protection against usual types of infection.

On the other hand, eosinophils are often produced in very large numbers in persons with parasitic infections, and they migrate into tissues diseased by parasites.

Though most of the parasites are too large to be phagocytized by the eosinophils or by any other phagocytic cells, nevertheless the eosinophils attach to the parasites and release substances that kill many of them. For instance, one of the most widespread infections of the world is *schistosomiasis*, a parasitic infection found in as many as one third of the population of some tropical countries, the parasite invades literally any part of the body. Eosinophils can attach to the juvenile forms of the parasite and can kill many of them. In our own country, another parasitic disease that causes eosinophilia is *trichinosis*, which results from invasion of the muscles by the *Trichinella* parasite ("pork worm") after a person eats uncooked pork.

Eosinophils also have a special propensity to collect in tissues in which allergic reactions have occurred, such as in the peribronchial tissues of the lungs in asthmatic persons, in the skin after allergic skin reactions, and so forth. This is caused at least partly by the fact that many mast cells and basophils participate in allergic reactions, as we shall discuss in the following section, and these cells release an *eosinophil chemotactic factor* that causes eosinophils to migrate toward the inflamed allergic tissue. The eosinophils are believed to detoxify some of the inflammation-inducing substances released by the mast cells and basophils and perhaps also by the damaged tissues, thus preventing spread of the local inflammatory process. For instance, eosinophils secrete substances that deactivate both heparin and histamine, which are among the collection of substances released by allergen-stimulated mast cells and basophils.

THE BASOPHILS

The basophils in the circulating blood are very similar to though not identical with the large *mast* cells located immediately outside many of the capillaries in the body. Both these cells liberate *heparin* into the blood, a substance that can prevent blood coagulation and that can also speed the removal of fat particles from the blood after a fatty meal.

The mast cells and basophils also release histamine, as well as smaller quantities of bradykinin and serotonin. Indeed, it is mainly the mast cells in inflamed tissues that release these substances during inflammation.

The mast cells and basophils play an exceedingly important role in some types of allergic reactions because the type of antibody that causes allergic reactions, the IgE type (see Chapter 7), has a special propensity to become attached to mast cells and basophils. Then, when the specific antigen subsequently reacts with the antibody, the resulting attachment of the antigen to the antibody causes the mast cell or basophil to rupture and release exceedingly large quantities of histamine, bradykinin, serotonin, heparin, and a number of lysosomal enzymes. These in turn cause local vascular and tissue reactions that cause the allergic manifestations. These effects will be discussed in greater detail in the following chapter.

AGRANULOCYTOSIS

A clinical condition known as "agranulocytosis" occasionally occurs, in which the bone marrow stops producing white blood cells, leaving the body unprotected against bacteria and other agents that might invade the tissues.

Normally, the human body lives in symbiosis with many bacteria, for all the mucous membranes of the body are constantly exposed to large numbers of bacteria. The mouth almost always contains various spirochetal, pneumococcal, and streptococcal bacteria, and these same bacteria are present to a lesser extent in the entire respiratory tract. The gastrointestinal tract is especially loaded with colon bacilli. Furthermore, one can almost always find bacteria in the eyes, the urethra, and the vagina. Therefore, any decrease in the number of neutrophils immediately allows invasion of the tissues by the bacteria that are already present in the body. Within two days after the bone marrow stops producing white blood cells, ulcers may appear in the mouth and colon, or the person develops some form of severe respiratory infection. Bacteria from the ulcers then rapidly invade the surrounding tissues and the blood. Without treatment, death often ensues three to six days after acute agranulocytosis begins.

Irradiation of the body by gamma rays caused by a nuclear explosion, or exposure to drugs and chemicals containing benzene or anthracene nuclei is quite likely to cause aplasia of the bone marrow. Indeed, some of the common drugs, such as chloramphenicol, thiouracil (used to treat thyrotoxicosis), and even the various barbiturate hypnotics, on occasion cause agranulocytosis (or *bone marrow aplasia* in which no cells of any type—red cells included—are produced in the bone marrow), thus setting off the entire infective sequence of this malady.

After irradiation injury to the bone marrow, a large number of stem cells, myeloblasts, and hemocytoblasts usually remain undestroyed and are capable of regenerating the bone marrow, provided sufficient time is available. Therefore, the patient properly treated with antibiotics and other drugs to ward off infection will usually develop enough new bone marrow within several weeks to several months that blood cell concentration can return to normal.

THE LEUKEMIAS

Uncontrolled production of white blood cells is caused by cancerous mutation of a myelogenous or a lymphogenous cell. This causes leukemia, which is usually characterized by greatly increased numbers of abnormal white blood cells in the circulating blood.

Types of Leukemia. Ordinarily, leukemias are divided into two general types: the *lymphogenous leukemias* and the *myelogenous leukemias*. The lymphogenous leukemias are caused by cancerous production of lymphoid cells, beginning first in a lymph node or other lymphogenous tissue and then spreading to other areas of the body. The second type of leukemia, myelogenous leukemia, begins by cancerous production of young myelogenous cells in the bone marrow and then spreads throughout the body so that white blood cells are produced in many extramedullary organs.

In myelogenous leukemia, the cancerous process occasionally produces partially differentiated cells, resulting in what might be called *neutrophilic leukemia*, *eosinophilic leukemia*, *basophilic leukemia*, or *monocytic leukemia*. More frequently, however, the leukemia cells

are bizarre and undifferentiated and not identical with any of the normal white blood cells. Usually the more undifferentiated the cells the more *acute* is the leukemia, often leading to death within a few months. But with some of the more differentiated cells, the process can be quite *chronic*, sometimes developing slowly over a period of 10 to 20 years.

Leukemic cells, especially the very undifferentiated cells, are usually nonfunctional, so that they cannot provide the usual protection associated with white blood cells.

EFFECTS OF LEUKEMIA ON THE BODY

The first effect of leukemia is metastatic growth of leukemic cells in abnormal areas of the body The leukemic cells of the bone marrow may reproduce so greatly that they invade the surrounding bone, causing pain and eventually a tendency to easy fracture. Almost all leukemias spread to the spleen, the lymph nodes, the liver, and other especially vascular regions, regardless of whether the origin of the leukemia is in the bone marrow or in the lymph nodes. In each of these areas the rapidly growing cells invade the surrounding tissues, utilizing the metabolic elements of these tissues and consequently causing tissue destruction.

Very common effects in leukemia are the development of infections, severe anemia, and bleeding tendency caused by thrombocytopenia (lack of platelets). These effects result mainly from displacement of the normal bone marrow by the leukemic cells.

Finally, perhaps the most important effect of leukemia on the body is the excessiva use of metabolic substrates by the growing cancerous cells The leukemic tissues reproduce new cells so rapidly that tremendous demands are made on the body fluids for foodstuffs, especially the amino acids and vitamins. Consequently, the energy of the patient is greatly depleted, and the excessive utilization of amino acids causes rapid deterioration of the normal protein tissues of the body. Thus, while the leukemic tissues grow, the other tissues are debilitated Obviously, after metabolic starvation has continued long enough, this alone is sufficient to cause death

REFERENCES

Allison, A C, *et al* Inflammation New York, Springer-Verlag, 1978

Askonas, B A, and Bancroft, G J Interaction of African trypanosomes with the immune system. *Phil Trans R Soc Lond [Biol], 307* 41, 1984

Barrett, K E, and Metcalfe, D D: Mast cell heterogeneity Evidence and implications *J Clin Immunol*, *4*:253, 1984

Braun, M. C Current Therapy in Hematology-Oncology 1983–1984 St. Louis, The C V Mosby Co, 1983

Escobar, M R, and Friedman, H (eds) Macrophages and Lymphocytes Nature, Functions and Interaction New York, Plenum Press, 1979

Ford-Hutchinson, A. W, Leukotrienes. Their formation and role as inflammatory mediators. *Fed Proc*, *44* 25, 1985

Foucar, K, *et al*: Nonendemic adult T-cell leukemia/lymphoma in the United States Report of two cases and review of the literature *Am J. Clin Pathol*, *83* 18, 1985

Freedman, M H. Congenital failure of hematopoiesis in the newborn infant *Clin. Perinatol*, *11* 417, 1984

Friedman, H, *et al* (eds.) The Reticuloendothelial System. New York, Plenum Press, 1979

Froese, A Receptors for IgE on mast cells and basophils. *Prog Allergy, 34* 142, 1984

Gabig, T G, and Babior, B M The killing of pathogens by phagocytes *Annu Rev Med*, *32* 313, 1981

Gadebusch, H J (ed) Phagocytosis and Cellular Immunity. West Palm Beach, Fla CRC Press, 1979

Galli, S J, *et al* Basophils and mast cells Morphologic insights into their biology, secretory patterns, and function *Prog Allergy, 34* 1, 1984

Goldstein, I. R M, and Marder, S R. Infections and hypocomplementemia. *Annu Rev Med*, *34* 47, 1983

Houck, J C (ed) Chemical Messengers of the Inflammatory Process. New York, Elsevier North Holland, 1980

Howard, J G, and Liew, F Y Mechanisms of acquired immunity in leishmaniasis. *Phil Trans R Soc Lond [Biol], 307* 87, 1984

Janoff A Neutrophil chemotaxis and mediation of tissue damage *In* Kaley, G, and Altura, B M (eds) Microcirculation Vol III Baltimore, University Park Press, 1977

Johannessen, J V Electron microscopy in human medicine *In* Cardiovascular System, Lymphoreticular and Hematopoietic System Vol. 5 New York, McGraw-Hill Book Co, 1982

Johnson, G R: Haemopoietic multipotential stem cells in culture *Clin Haematol*, *13* 309, 1984

Keleman, E, *et al* Atlas of Human Hemopoietic Development. New York, Springer-Verlag, 1979

Klebanoff, S J, and Clark, R A: The Neutrophil Function and Clinical Disorders New York, Elsevier/North-Holland, 1978.

Kokubun, Y, and Kobayashi, N (eds) Phagocytosis, Its Physiology and Pathology Baltimore, University Park Press, 1979

Lagunoff, D, *et al* Agents that release histamine from mast cells. *Annu. Rev Pharmacol Toxicol*, *23* 331, 1983

Lichtman M A (ed) Hematology and Oncology New York, Grune & Stratton, 1980

Lisiewicz, J Human Neutrophils Bowie, Md, Charles Press Publishers, 1979

Mackinney, A. A, Jr The Pathophysiology of Blood New York, John Wiley & Sons, 1984

Mahapatra D, and Mahapatra, R C Cutaneous basophil hypersensitivity *Am J Dermatopathol*, *6* 483, 1984

May, M E, and Waddell, C C Basophils in peripheral blood and bone marrow A retrospective review *Am J Med*, *76* 509, 1984

Ogawa, M, *et al* Differentiation and proliferative kinetics of hemopoietic stem cells in culture *Prog Clin Biol Res*, *143* 35, 1984

Olsen, E G, and Spry, C J Relation between eosinophilia and endomyocardial disease *Prog Cardiovasc Dis*, *27* 241, 1985

Peterson, P K, and Quie, P G Bacterial surface components and the pathogenesis of infectious diseases *Annu Rev Med*, *32* 29, 1981.

Quastel, M R (ed) Cell Biology and Immunology of Leukocyte Function New York, Academic Press 1979

Rebuck, J W, *et al* (eds) The Reticuloendothelial System. Huntington, N Y, R E Krieger Publishing Co, 1980

Rowley, J D Biological implications of consistent chromosome rearrangements in leukemia and lymphoma *Cancer Res*, *44* 3159, 1984

Seiverd, C E Hematology for Medical Technologists. Philadelphia, Lea & Febiger, 1983

Shiffmann, E Leukocyte chemotaxis *Annu Rev Physiol*, *44* 553, 1982.

Simone, J V Outlook for acute lymphocytic leukemia in children in 1982 *Annu Rev Med*, *32* 207, 1981

Spivak, J L (ed) Fundamentals of Clinical Hematology Hagerstown, Md., Harper & Row, 1980

Stanley, E R, and Jubinsky, P T Factors affecting the growth and differentiation of haemopoietic cells in culture *Clin Haematol*, *13* 329, 1984

Thompson, R. A A Short Textbook of Haematology Baltimore, Urban & Schwarzenberg, 1984

Van Furth, R. (ed.) Mononuclear Phagocytes Functional Aspects of Mononuclear Phagocytes. Boston, M. Nijhoff, 1979

Verhoef, J, and Verbrugh, H A Host determinants in staphylococcal disease *Annu Rev. Med*, *32* 107, 1981.

Warnke, R. A, and Link, M P. Identification and significance of cell markers in leukemia and lymphoma *Annu Rev Med*, *34* 117, 1983

6

Immunity and Allergy

INNATE IMMUNITY

The human body has the ability to resist almost all types of organisms or toxins that tend to damage the tissues and organs. This capacity is called *immunity*. Much of the immunity is caused by a special immune system that forms antibodies and activated lymphocytes that attack and destroy the specific organisms or toxins. This type of immunity is *acquired immunity*. However, an additional portion of the immunity results from general processes rather than from processes directed at specific disease organisms. This is called *innate immunity*. It includes the following:

1. Phagocytosis of bacteria and other invaders by white blood cells and cells of the tissue macrophage system, as described in the previous chapter.
2. Destruction by the acid secretions of the stomach and by the digestive enzymes of organisms swallowed into the stomach.
3. Resistance of the skin to invasion by organisms.
4. Presence in the blood of certain chemical compounds that attach to foreign organisms or toxins and destroy them. Some of these are (a) *lysozyme*, a mucolytic polysaccharide that attacks bacteria and causes them to dissolute, (b) *basic polypeptides*, which react with and inactivate certain types of gram-positive bacteria; and (c) the *complement complex* that will be described later, a system of about 20 proteins that can be activated in various ways to destroy bacteria.

This innate immunity makes the human body resistant to such diseases as some paralytic virus infections of animals, hog cholera, cattle plague, and distemper—a viral disease that kills a large percentage of dogs that become afflicted with it. On the other hand, lower animals are resistant or even completely immune to many human diseases, such as poliomyelitis, mumps, human cholera, measles, and syphilis, which are very destructive or even lethal to the human being.

ACQUIRED IMMUNITY

In addition to its innate immunity, the human body also has the ability to develop extremely powerful specific immunity against individual invading agents such as lethal bacteria, viruses, toxins, and even foreign tissues from other animals. This is called *acquired immunity*. It is with this immune mechanism and some of its associated reactions—especially the allergies—that the remainder of this chapter will be concerned.

Acquired immunity can often bestow extreme protection. For instance, certain toxins such as the paralytic toxin of botulinum or the tetanizing toxin of tetanus can be protected against in doses as high as 100,000 times the amount that would be lethal without immunity. This is the reason the process known as "vaccination" is so extremely important in protecting human beings against disease and against toxins, as will be explained in the course of this chapter.

TWO BASIC TYPES OF ACQUIRED IMMUNITY

Two basic, but closely allied, types of acquired immunity occur in the body. In one of these the body develops circulating *antibodies*, which are globulin molecules that are capable of attacking the invading agent. This type of immunity is called *humoral immunity*. The second type of acquired immunity is achieved through the formation of large numbers of *activated lymphocytes* that are specifically designed to destroy the foreign agent. This type of immunity is called *cell-mediated immunity*.

We shall see shortly that both the antibodies and the activated lymphocytes are formed in the lymphoid tissue of the body. First, let us discuss the initiation of the immune process by *antigens*.

ANTIGENS

Since acquired immunity does not occur until after first invasion by a foreign organism or toxin, it is clear that the body must have some mechanism for recognizing the initial invasion. Each toxin or each type of organism almost always contains one or more specific chemical compounds in its makeup that are different from all other compounds. In general, these are proteins or large polysaccharides, and it is they that cause the acquired immunity. These substances are called *antigens*.

For a substance to be antigenic it usually must have a high molecular weight, 8000 or greater. Furthermore, the process of antigenicity usually depends upon regularly recurring prosthetic radicals on the surface of the large molecule, which perhaps explains why proteins and polysaccharides are almost always antigenic, for they both have this type of stereochemical characteristic

Haptens. Though substances with molecular weights of less than 8000 only rarely act as antigens, immunity can nevertheless be developed against substances of low molecular weight in a very special way, as follows: If the low molecular weight compound, which is called a *hapten*, first combines with a substance that is antigenic, such as a protein, then the combination will elicit an immune response. The antibodies or activated lymphocytes that develop against the combination can then react either against the protein or against the hapten. Therefore, on second exposure to the hapten, some of the antibodies or lymphocytes react with it before it can spread through the body and cause damage.

The haptens that elicit immune responses of this type are usually drugs, chemical constituents in dust, breakdown products of dandruff from animals, degenerative products of scaling skin, various industrial chemicals, the toxin of poison ivy, and so forth.

ROLE OF LYMPHOID TISSUE IN ACQUIRED IMMUNITY

Acquired immunity is the product of the body's lymphoid tissue. In persons who have a genetic lack of lymphoid tissue or whose lymphoid tissue has been destroyed by radiation or by chemicals, no acquired immunity whatsoever can develop. And almost immediately after birth such a person dies of fulminating infection unless treated by heroic measures. Therefore, it is clear that the lymphoid tissue is essential to survival of the human being.

The lymphoid tissue is located most extensively in the *lymph nodes*, but it is also found in special lymphoid tissues such as that of the *spleen*, *submucosal areas of the gastrointestinal tract*, and the *bone marrow*. The lymphoid tissue is distributed very advantageously in the body to intercept the invading organisms or toxins before they can spread too widely. For instance, the lymphoid tissue of the gastrointestinal tract is exposed immediately to antigens invading through the gut. The lymphoid tissue of the throat and pharynx (the tonsils and adenoids) is extremely well located to intercept antigens that enter by way of the upper respiratory tract. The lymphoid tissue in the lymph nodes is exposed to antigens that invade the peripheral tissues of the body. And, finally, the lymphoid tissue of the spleen and bone marrow plays the specific role of intercepting antigenic agents that have succeeded in reaching the circulating blood.

Two Types of Lymphocytes That Promote, Respectively, Cell-Mediated Immunity and Humoral Immunity—The T and the B Lymphocytes. Though most of the lymphocytes in normal lymphoid tissue look alike when studied under the microscope, these cells are distinctly divided into two major populations. One of the populations is responsible for forming the activated lymphocytes that provide cell-mediated immunity and the other for forming the antibodies that provide humoral immunity.

Both of these types of lymphocytes are derived originally in the embryo from *pluripotent hemopoietic stem cells* that differentiate and become committed to form lymphocytes. The lymphocytes that are formed eventually end up in the lymphoid tissue, but before doing so they are further differentiated or "preprocessed" in the following ways:

Those lymphocytes that are eventually destined to form activated lymphocytes first migrate to and are preprocessed in the *thymus* gland, for which reason they are called T lymphocytes. These are responsible for cell-mediated immunity.

The other population of lymphocytes—those that are destined to form antibodies—are preprocessed in some unknown area of the body, probably in the liver during midfetal life and in the bone marrow in late fetal life and after birth. However, this population of cells was first discovered in birds in which the preprocessing occurs in the *bursa of Fabricius*, a structure not found in mammals. For this reason these lymphocytes are called B lymphocytes, and they are responsible for humoral immunity.

Figure 6–1 illustrates the two separate lymphocyte systems for the formation, respectively, of the activated T lymphocytes and the antibodies.

PREPROCESSING OF THE T AND B LYMPHOCYTES

Though all the lymphocytes of the body originate from the lymphocytic committed stem cells of the embryo,

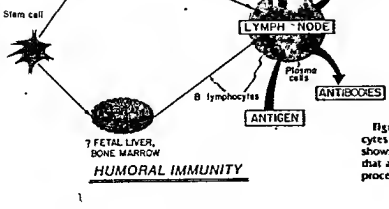


Figure 6–1. Formation of antibodies and sensitized lymphocytes by a lymph node in response to antigens. This figure shows the origin of *thymic* ("T") and *bursal* ("B") that are responsible for the cell-mediated and humoral processes of the lymph nodes.

these stem cells are themselves incapable of forming either activated lymphocytes or antibodies. Before they can do so, they must be further differentiated in appropriate processing areas in the thymus or in the B cell processing area

Role of the Thymus Gland for Preprocessing the T Lymphocytes. Most of the preprocessing of the T lymphocytes of the thymus gland occurs shortly before the birth of the baby and for a few months after birth. Therefore, beyond this period of time, removal of the thymus gland usually will not seriously impair the T lymphocytic immune system, the system necessary for cell-mediated immunity. However, removal of the thymus several months before birth can completely prevent development of all cell-mediated immunity. Since it is this cellular type of immunity that is mainly responsible for rejection of transplanted organs such as hearts and kidneys, one can transplant organs with little likelihood of rejection if the thymus is removed from an animal a reasonable period of time before birth.

Thymic Hormone. In addition to preprocessing the T lymphocytes, the thymus probably also secretes a hormone, called *thymopoietin*, that circulates through the body fluids and increases the activity of the T lymphocytes that have already left the thymus gland and have migrated to the lymphoid tissue. This hormone is believed to cause further proliferation and increased activity of these lymphocytes Otherwise, little is known about either the nature or the function of this hormone

Role of the Bursa of Fabricius for Preprocessing B Lymphocytes in Birds. During the latter part of fetal life the bursa of Fabricius preprocesses the B lymphocytes and prepares them to manufacture antibodies. Here again, this preprocessing continues for a while after birth. In mammals, it is believed that the B cells are preprocessed during midfetal life in the liver but thereafter in the bone marrow.

Spread of Processed Lymphocytes to the Lymphoid Tissue. After formation of processed lymphocytes in both the thymus and the bursa, these first circulate freely in the blood for a few hours but then become entrapped in the lymphoid tissue Thus, the lymphocytes do not originate primordially in the lymphoid tissue but, instead, are transported to this tissue by way of the preprocessing areas of the thymus and probably fetal liver and bone marrow

SPECIFICITY OF ANTIBODIES AND T LYMPHOCYTES— ROLE OF LYMPHOCYTE CLONES

When either antibodies or activated T lymphocytes, called "T cells," are formed in the lymphoid tissue, these then can react highly specifically against the particular type of antigen that initiated their development. Formerly, it was believed that this was achieved by the antigen's in some way modifying the lymphocytes in the lymph node and in this way causing them to form either antibodies or T cells with specificity for that particular type of antigen However, it is now known that this is not true Instead the mechanism is almost certainly the following

Millions of Preformed Specific Lymphocytes in the Lymphoid Tissue. We now know that there are at least a million different types of preformed B lymphocytes and equally as many preformed T lymphocytes that are capable of forming the highly specific antibodies, or T cells, when activated by the appropriate antigens. Each one of these preformed lymphocytes is capable of forming only one type of antibody or one type of T cell with a single type of specificity. And only the specific type of antigen with which it can react can activate it. However, once the specific lymphocyte is activated by its antigen, it reproduces wildly, forming tremendous numbers of duplicate lymphocytes. If it is a B lymphocyte, its progeny will eventually secrete antibodies that then circulate throughout the body. If it is a T lymphocyte, its progeny are sensitized T cells that are released into the lymph and carried to the blood, then circulated through all the tissue fluids and back into the lymph, circulating around and around in this circuit sometimes for months or years.

All the different lymphocytes that are capable of forming one specificity of antibody or T cell are called a *clone of lymphocytes*. That is, the lymphocytes in each clone are alike and are derived originally from one or a few early lymphocytes of its specific type.

ORIGIN OF THE MANY CLONES OF LYMPHOCYTES

In lower animals, and presumably also in human beings, only about a thousand genes code for the different types of antibodies and T lymphocytes. At first, it was a mystery how it was possible for so few genes to code for the million or more different specificities of antibody molecules or T cells that can be produced by the lymphoid tissue, especially when one considers that a single gene is usually necessary for the formation of each different type of protein However, the answer to this has now been discovered. During the preprocessing stage, the genes cause the formation of large RNA molecules containing many small functional segments, each of which codes for an individual portion of an antibody or of a T cell "marker" (the protein on the T cell surface that gives it specificity), but not for the entire antibody or T cell marker. Then, while still in the nucleus, the RNA molecule is cut into its segments and respliced to form the final messenger RNA molecule, employing the cutting and splicing mechanism that was described in Chapter 3 However, during this cutting and splicing the individual segments of the original RNA molecule are spliced in different combinations in different ones of the processed lymphocytes And when one realizes that there are a thousand or more different genes, each coding for multiple segments of antibodies or cell markers, one can understand what a vast number of different combinations can be formed.

Mechanism for Activating a Clone of Lymphocytes

Each clone of lymphocytes is responsive to only a single type of antigen (or to similar antigens that have almost exactly the same stereochemical characteristics) The reason for this is the following In the case of the B lymphocytes, each one of them has in its cell membrane about 100,000 antibody molecules that will react highly specifically with only the one specific type of antigen Therefore, when the appropriate antigen comes along, it immediately attaches to the cell membrane; this leads to the activation process, which we will describe in more detail subsequently. In the case of the T lymphocytes, molecules very similar to antibodies, called *surface re-*

ceptor proteins (or T cell markers), are in the cell membrane, and these too are highly specific for the one specified activating antigen.

Role of Macrophages in the Activation Process. Aside from the lymphocytes in lymphoid tissue, literally millions of macrophages are also present in the same tissue. These line the sinusoids of the lymph nodes, spleen, and other lymphoid tissue, and they lie in apposition to many of the lymph node lymphocytes. Most invading organisms are first phagocytized and partially digested by the macrophages, and the antigenic products are liberated into the macrophage cytosol. It is believed that the macrophages then pass these antigens directly to the lymphocytes, thus leading to activation of the specified clones. The macrophages also secrete an activating substance that promotes the growth and reproduction of the specific lymphocytes. This substance is called *interleukin-1*.

Role of the T cells in the Activation of the B Lymphocytes. Most antigens activate both T lymphocytes and B lymphocytes at the same time. Some of the T cells that are formed, called *helper cells*, in turn secrete specific substances (collectively called "lymphokines") that further activate the B lymphocytes. Indeed, without the help of these T cells, the quantity of antibodies formed by the B lymphocytes is usually very slight. Two of the lymphokines secreted by the helper cells are (1) *B cell growth factor* also called *interleukin-2*, that causes growth and reproduction of the B cells to form the intracellular organelles needed to produce antibodies. We shall discuss this cooperative relationship between helper cells and B cells again after we have had an opportunity to describe the mechanisms of the T cell system of immunity.

SPECIFIC ATTRIBUTES OF THE B LYMPHOCYTE SYSTEM—HUMORAL IMMUNITY AND THE ANTIBODIES

Formation of Antibodies by the Plasma Cells. Prior to exposure to a specific antigen, the clones of B lymphocytes remain dormant in the lymphoid tissue. However, upon entry of a foreign antigen, the lymphoid tissue macrophages phagocytize the antigen and then present it to the adjacent B lymphocytes. In addition, the antigen may also be presented to T cells at the same time, and one type of these, the "helper" T cells, then also contributes to the activation of the B lymphocytes, as we shall discuss more fully later. Those B lymphocytes specific for the antigen immediately enlarge and take on the appearance of *lymphoblasts*. Some of these then further differentiate to form *plasmablasts*, which are the precursors of *plasma cells*. In these cells the cytoplasm expands, and the rough endoplasmic reticulum vastly proliferates. They then begin to divide at a rate of approximately once every ten hours for about nine divisions, giving in four days a total population of about 500 cells for each original plasmablast. The mature plasma cell then produces gamma globulin antibodies at an extremely rapid rate—about 2000 molecules per second for each plasma cell. The antibodies are secreted into the lymph and are carried to the circulating blood. This process continues for several days or weeks until death of the plasma cells.

Formation of "Memory" Cells—Difference Between the Primary Response and the Secondary Response. Some of the lymphoblasts formed by activation of a clone of B lymphocytes do not go on to form plasma cells but, instead, form moderate numbers of new B lymphocytes similar to those of the original clone. In other words, the population of the specifically activated clone becomes greatly enhanced. And the new B lymphocytes are added to the original lymphocytes of the clone. They also circulate throughout the body to inhabit all the lymphoid tissue, but immunologically they remain dormant until activated once again by a new quantity of the same antigen. These are called *memory cells*. Obviously, subsequent exposure to the same antigen will then cause a much more rapid and much more potent antibody response.

Figure 6–2 illustrates the differences between the *primary response* that occurs on first exposure to a specific antigen and the *secondary response* that occurs following a second exposure to the same antigen. Note the delay in the appearance of the primary response, its weak potency, and its short life. The secondary response, by contrast, begins rapidly after exposure to the antigen, is far more potent, and forms antibodies for many months rather than for only a few weeks.

The increased potency and duration of the secondary response explains why *vaccination* is usually accomplished by injecting antigen in multiple doses with periods of several weeks or several months between injections.

The Nature of the Antibodies

The antibodies are gamma globulins called *immunoglobulins*, and they have molecular weights between approximately 150,000 and 900,000. Usually they constitute about 20 per cent of all the plasma proteins.

All the immunoglobulins are composed of combinations of *light* and *heavy polypeptide chains*, most of which are a combination of two light and two heavy chains, as illustrated in Figure 6–3. Some of the immunoglobulins, though, have combinations of as many as ten heavy and ten light chains, which gives rise to the much larger molecular weight immunoglobulins. Yet, in all immunoglobulins, each heavy chain is paralleled by a light chain at one of its ends, thus forming a heavy-light pair, and there are always at least two such pairs in each immunoglobulin molecule.

Figure 6–3 shows a designated end of each of the light and each of the heavy chains, called the "variable portion," and the remainder of each chain is called the

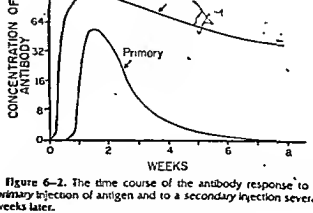


Figure 6–2. The time course of the antibody response to a *primary* injection of antigen and to a *secondary* injection several weeks later.

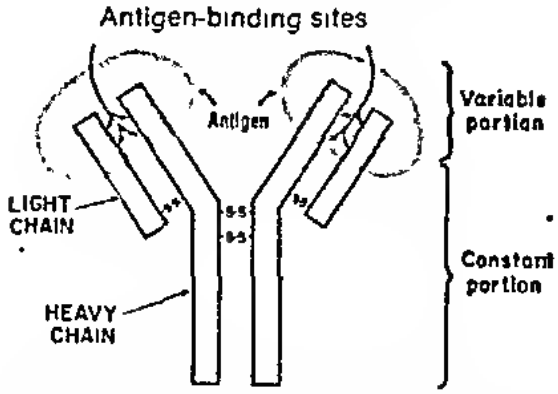


Figure 6–3. Structure of the typical IgG antibody, showing it to be composed of two heavy polypeptide chains and two light polypeptide chains. The antigen binds at two different sites on the variable portions of the chains.

"constant portion." The variable portion is different for each specificity of antibody, and it is this portion that attaches specifically to a particular type of antigen. The constant portion of the antibody determines much of the gross physical and chemical properties of the antibody, establishing such factors as diffusivity of the antibody in the tissues, adherence of the antibody to specific structures within the tissues, attachment to the *complement complex*, the ease with which the antibodies pass through membranes, and other biological properties of the antibody.

Specificity of Antibodies. Each antibody that is specific for a particular antigen has a different organization of amino acid residues in the variable portions of both the light and heavy chains. These have a different steric shape for each antigen specificity, so that when an antigen comes in contact with it the prosthetic radicals of the antigen fit as a mirror image with those of the antibody, thus allowing a rapid bonding between the antibody and the antigen. The bonding is noncovalent, but when the antibody is highly specific there are so many bonding sites that the antibody-antigen coupling is nevertheless exceedingly strong, held together by (a) hydrophobic bonding, (b) hydrogen bonding, (c) ionic attractions, and (d) van der Waals forces. It also obeys the thermodynamic mass action law

$$K_a = \frac{\text{concentration of bound antibody-antigen}}{\text{concentration of antibody} \times \text{concentration of antigen}}$$

K_a is called the *affinity constant* and is a measure of how tightly the antibody binds with the antigen.

Note, especially, in Figure 6–3 that there are two variable sites on the antibody for attachment of antigens. Thus, most antibodies are *bivalent*. However, a small proportion of the antibodies, which have high molecular-weight combinations of light and heavy chains, have as many as ten *binding sites*.

Classes of Antibodies. There are five general classes of antibodies, respectively named *IgM, IgG, IgA, IgD,* and *IgE*. Ig stands for immunoglobulin, and the other five respective letters simply designate the respective classes.

For the purpose of our present limited discussion, two of these classes of antibodies are of particular importance: IgG, which comprises about 75 per cent of the antibodies of the normal person, and IgE, which constitutes only a small per cent of the antibodies but which is especially involved in allergy. However, the IgM class is also interesting because a large share of the antibodies formed during the primary response are of this type. They have ten binding sites, which makes them exceedingly effective in protecting the body against invaders even though there are not many IgM antibodies.

Mechanisms of Action of Antibodies

Antibodies act mainly in two different ways to protect the body against invading agents: (1) by direct attack on the invader, or (2) by activation of the complement system that then destroys the invader.

Direct Action of Antibodies on Invading Agents. Figure 6–4 illustrates antibodies (designated by the bars) reacting with antigens (designated by the darkened dumbbells). Because of the bivalent nature of the antibodies and the multiple antigen sites on most invading agents, the antibodies can inactivate the invading agent in one of several ways, as follows.

1. *Agglutination*, in which multiple large structures with antigens on their surfaces, such as bacteria or red cells, are bound together into a clump.
2. *Precipitation*, in which the molecular complex of soluble antigen (such as tetanus toxin) and antibody becomes so large that it is rendered insoluble and precipitates.
3. *Neutralization*, in which the antibodies cover the toxic sites of the antigenic agent.
4. *Lysis*, in which some very potent antibodies are occasionally capable of directly attacking membranes of cellular agents and thereby causing rupture of the cell.

However, the direct actions of antibodies attacking the antigenic invaders probably, under normal conditions, are not strong enough to play a major role in protecting the body against the invader. Most of the protection comes through the *amplifying* effects of the complement system described below.

The Complement System for Antibody Action. "Complement" is a collective term to describe a system of about 20 different proteins, many of which are enzyme precursors. The principal actors in this system are 11 proteins designated C1 through C9, B, and D, illustrated in Figure 6–5. All these are present normally among the plasma proteins and also among the plasma proteins that leak out of the capillaries into the tissue spaces. The enzyme precursors are normally inactive, but they can be activated in two separate ways called (a) the *classical pathway* and (b) the *alternate pathway*.

The Classical Pathway. The classical pathway is activated by an antigen-antibody reaction. That is, when

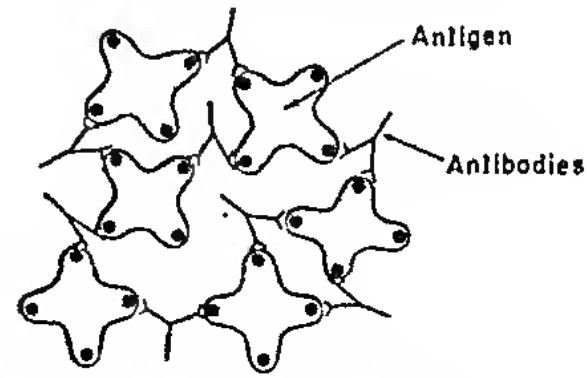


Figure 6–4. Binding of antigen molecules to each other by bivalent antibodies.

an antibody binds with an antigen, a specific reactive site on the "constant" portion of the antibody becomes uncovered, or activated, and this in turn binds directly with the C1 molecule of the complement system, setting into motion a "cascade" of sequential reactions, illustrated in Figure 6–5, beginning with activation of the proenzyme C1 itself. Only a few antigen-antibody combinations are required to activate many molecules in this first stage of the complement system. The C1 enzymes thus formed then activate successively increasing quantities of enzymes in the later stages of the system, so that from a very small beginning a very large reaction occurs. Multiple end-products are formed as illustrated in the figure, and several of these cause important effects that help to prevent damage by the invading organism or toxin. Among the more important effects are the following:

(1) *Opsonization* and *phagocytosis*. One of the products of the complement cascade, C3b, strongly activates phagocytosis by both neutrophils and macrophages, causing them to engulf the bacteria to which the antigen-antibody complexes are attached. This process is called *opsonization*. It often enhances the number of bacteria that can be destroyed by many hundredfold.

(2) *Lysis*. One of the most important of all the products of the complement cascade is the *lytic complex*, which is a combination of multiple complement factors and is designated C5b6789. This has a direct effect of rupturing the cell membranes of bacteria or other invading organisms.

(3) *Agglutination*. The complement products also change the surfaces of the invading organisms, causing them to adhere to each other, thus promoting agglutination.

(4) *Neutralization of viruses*. The complement enzymes and other complement products can attack the structures of some viruses and thereby render them nonvirulent.

(5) *Chemotaxis*. Fragment C5a causes chemotaxis of neutrophils and macrophages, thus causing large numbers of these phagocytes to migrate into the local region of the antigenic agent.

(6) *Activation of mast cells and basophils*. Fragments C3a, C4a, and C5a all activate mast cells and basophils, causing them to release histamine and several other substances into the local fluids. These substances in turn cause increased local blood flow, increased leakage of fluid and plasma protein into the tissue, and other local tissue reactions that help inactivate or immobilize the antigenic agent. These same factors play a major role in inflammation, which was discussed in the preceding chapter, and also in allergy, as we shall discuss later.

(7) *Inflammatory effects*. In addition to inflammatory effects caused by activation of the mast cells and basophils, several other complement products contribute further to local inflammation. These products cause the hyperemia to be increased, the leakage of proteins to be increased, and the proteins to coagulate in the tissue spaces, thus preventing movement of the invading organism through the tissues.

The Alternate Pathway. The complement system can sometimes be activated without the intermediation of an antigen-antibody reaction. This occurs especially in response to large polysaccharide molecules in the cell membranes of some invading microorganisms. These substances react with complement factors B and D, forming an activation product that activates factor C3, setting off the remainder of the complement cascade beyond the C3 level. Thus, essentially all the same final products of the system are formed as in the classical pathway, and these cause the same effects as those just listed in protecting the body against the invader.

Since the alternate pathway does not involve an antigen-antibody reaction, it is one of the first lines of defense against invading microorganisms, capable of functioning even before a person becomes immunized against the organism.

SPECIAL ATTRIBUTES OF THE T LYMPHOCYTE SYSTEM—ACTIVATED CELLS AND "CELL-MEDIATED IMMUNITY"

Release of Activated T Cells from Lymphoid Tissue, and Formation of Memory Cells. Upon exposure to the proper antigens, the T lymphocytes of the lymphoid tissue proliferate and release large numbers of activated T cells in ways that parallel antibody release. The only real difference is that instead of releasing

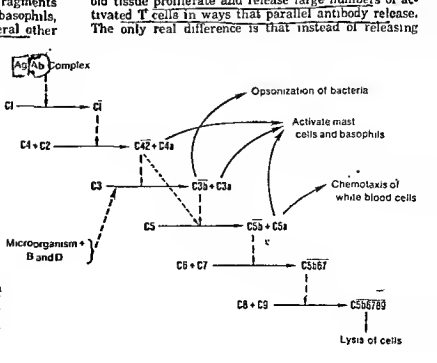


Figure 6–5. Cascade of reactions during activation of the classical pathway of complement. (Modified from Alexander and Good Fundamentals of Clinical Immunology. Philadelphia, W. B. Saunders Company, 1977.)

antibodies, whole activated T cells are formed and released into the lymph. These then pass into the circulation and are distributed throughout the body, passing through the capillary walls into the tissue spaces, back into the lymph and blood once again, and circulating again and again throughout the body, sometimes lasting for months or even years.

Also, *T lymphocyte memory cells* are formed in the same way that memory cells are formed in the humoral antibody system. That is, when a clone of T lymphocytes is activated by an antigen, many of the newly formed lymphocytes are preserved in the lymphoid tissue to become additional T lymphocytes of that specific clone; in fact, these memory cells even spread throughout the lymphoid tissue of the entire body. Therefore, on subsequent exposure to the same antigen, the release of activated T cells occurs far more rapidly and much more powerfully than in the first response.

Antigen Receptors on the T Lymphocytes. Antigens bind with *receptor molecules* on the surfaces of T cells in the same way that they bind with antibodies. These receptor molecules are composed of a variable unit similar to the variable portion of the humoral antibody, but its stem section is firmly bound to the cell membrane. Therefore, it is never secreted frm the cell into the body fluids. There are as many as 100,000 receptor sites on a single T cell.

Multiple Types of T Cells and Their Functions

In the last few years it has become clear that there are several different types of T cells. These are classified into three major groups: (1) *cytotoxic T cells*, (2) *helper T cells*, and (3) *suppressor T cells*. The functions of each of these are quite distinct.

Cytotoxic T Cells. The cytotoxic T cell is a direct attack cell, capable of killing microorganisms and at times even some of the body's own cells. For this reason, these cells are frequently called *killer cells*. The receptor proteins on the surfaces of the cytotoxic cells cause them to bind tightly to those organisms or cells that contain their binding-specific antigen. Then, they kill the attacked cell in the manner illustrated in Figure 6–6. The immediate effect of the binding is swelling of the T cell and release of cytotoxic substances directly into the attacked cell. The cytotoxic substances are probably mainly lysosomal enzymes manufactured in the T cell. It is especially important that each of these "killer" cells can attack many different microorganisms one after another, killing many separate organisms instead of merely one, often without the killer cells themselves being harmed.

The cytotoxic T cells are especially lethal to tissue cells that have been invaded by viruses because many virus particles become entrapped in the membranes of these cells and attract the T cells because of the viral antigenicity. The cytotoxic cells also play an important role in destroying cancer cells, heart transplant cells, or other types of cells that are "foreign" to the person's own body.

The Helper T Cells. The helper T cells are by far the most numerous of the T cells. As their name implies they "help" in the functions of the immune system in multiple ways. Some of their important functions are:

(1) *Increasing the activation of B cells, cytotoxic T cells, and suppressor T cells by antigens.* The B lymphocyte clones in the lymphoid tissue, as well as the cytotoxic T cells and suppressor T cell clones, are usually activated poorly by most antigens in the absence of helper T cells. In contrast, the helper T cell clones can be activated by very small quantities of antigens and even by antigens that have very few reactive sites whereas the other cells require many sites for direct activation. However, once the helper T cells have been activated, they in turn secrete "lymphokines" that enhance the response of the other types of lymphoid cells to the antigen, as was explained earlier.

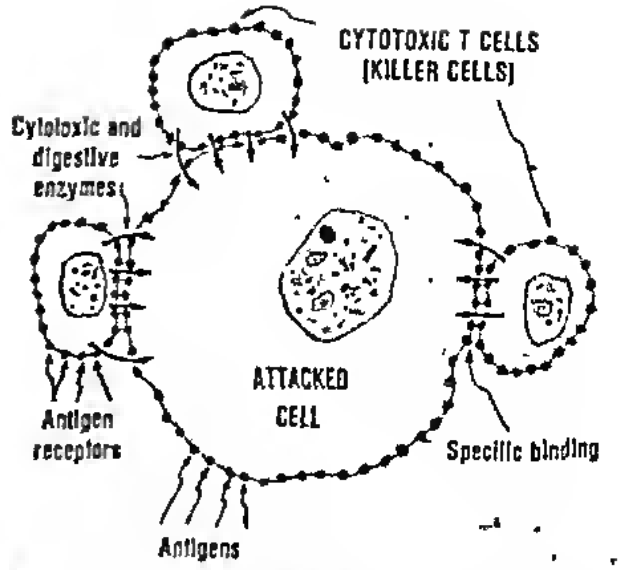


Figure 6–6. Direct destruction of an invading cell by sensitized lymphocytes.

(2) *Stimulation of activity of other T cells.* The helper T cells secrete a substance called *interleukin-2*, one of the lymphokines, that increases the activity of other T cells, including cytotoxic T cells, some suppressor T cells, and probably even some of the other helper T cells.

(3) *Activation of the macrophage system.* The helper T cells secrete another lymphokine, called *macrophage migration inhibition factor*, that plays two important roles in the macrophage response to invading organisms. First, it slows or stops the migration of macrophages that have been chemotactically attracted into the infected tissue area, thus causing a great accumulation of macrophages. Second, it activates the macrophages to cause far more efficient phagocytosis, allowing them to attack and destroy greatly increased numbers of invading organisms.

Suppressor T Cells. Much less is known about the suppressor T cells than about the others, but they are capable of suppressing the functions of both cytotoxic and helper T cells. It is believed that these suppressor functions serve the purpose of regulating the activities of the other cells, keeping them from causing excessive immune reactions that might be severely damaging to the body. For this reason, the suppressor cells are frequently called *regulatory T cells*. One scenario for the function of the regulatory suppressor T cells is the following: the helper T cells activate the suppressor T cells; these cells in turn act as a negative feedback controller of the helper T cells; and this automatically sets the level of activity of the helper T cell system. It is also probable that the suppressor T cells play an important role in limiting the ability of the immune system to attack a person's own body tissues, called *immune tolerance*, as we shall discuss in the following section.

TOLERANCE OF THE ACQUIRED IMMUNITY SYSTEM TO ONE'S OWN TISSUES—ROLE OF THE THYMUS AND THE BURSA

Obviously, if a person should become immune to his or her own tissues, the process of acquired immunity would destroy the individual's own body. Fortunately, the immune mechanism normally "recognizes" a person's own tissues as being completely distinctive from those of invaders, and his immunity system forms very few antibodies or activated T cells against his own antigens. This phenomenon is known as *tolerance* to the body's own tissues.

Mechanism of Tolerance. It is believed that most of the phenomenon of tolerance develops during the processing of the T lymphocytes in the thymus and the B lymphocytes in the B lymphocyte processing area—in birds in the bursa of Fabricius. The reason for this belief is that injecting a strong antigen into a fetus at the time that the lymphocytes are being processed in these two areas will prevent the development of clones of lymphocytes in the lymphoid tissue that are specific for the injected antigen. Also, experiments have shown that specific immature lymphocytes in the thymus, when exposed to a strong antigen, become lymphoblastic, proliferate considerably, and then combine with the stimulating antigen—an effect that is believed to cause the cells themselves to be destroyed before they can migrate to and colonize the lymphoid tissue.

Therefore, it is believed that during the processing of lymphocytes in the thymus and in the B lymphocyte processing area, all those clones of lymphocytes that are specific for the body's own tissues are self-destroyed because of their continual exposure to the body's antigens.

However, still another method by which tolerance might occur has recently been suggested, based on the suppressor function of the suppressor T cells. In a few instances it has been noted that increase in the number of suppressor cells is associated with reduced immunity against some antigens. Therefore, it is theoretically possible that autoimmunity against the body's own antigens might be suppressed by strong development of specific suppressor T cells; as yet, this is not entirely clear.

Failure of the Tolerance Mechanism—Autoimmune Diseases. Unfortunately, people frequently lose some of their immune tolerance to their own tissues. This occurs to a greater extent the older a person becomes. It usually occurs after destruction of some of the body's tissues, which releases considerable quantities of antigens that circulate in the body and presumably cause acquired immunity in the form of either activated T cells or antibodies. Some of these antigens seem to combine with other proteins, such as proteins from bacteria or viruses, to form a new type of antigen that can then cause immunity. Then the resulting activated T cells and antibodies attack the body's own tissues. Also, it is believed that some of the proteins of the body are normally sequestered from the immune system during embryonic development of tolerance so that tolerance to these proteins never develops in the first place. For instance, the proteins of the cornea do not seem to circulate in the fluids of the fetus; this is also true of the thyroglobulin molecule of the thyroid; therefore, tolerance to these never develops. When damage occurs to either of these two tissues, these protein molecules can then elicit immunity, and the immunity in turn can attack the cornea in the first instance or the thyroid gland in the second instance to cause corneal opacity or thyroiditis.

Other diseases that result from autoimmunity include *rheumatic fever*, in which the body becomes immunized against tissues in the joints and heart, especially the heart valves, following exposure to a specific type of streptococcal toxin; one type of *glomerulonephritis*, in which the person becomes immunized against the basement membranes of glomeruli; *myasthenia gravis*, in which immunity develops against the acetylcholine receptor proteins of the neuromuscular junction, causing paralysis; and *lupus erythematosus*, in which the person becomes immunized against many different body tissues at the same time, a disease that causes extensive damage and often rapid death.

VACCINATION

The process of vaccination has been used for many years to cause acquired immunity against specific diseases. A person can be vaccinated by injecting dead organisms that are no longer capable of causing disease but which still have their chemical antigens. This type of vaccination is used to protect against typhoid fever, whooping cough, diphtheria, and many other types of bacterial diseases. Also, immunity can be achieved against toxins that have been treated with chemicals so that their toxic nature has been destroyed even though their antigens for causing immunity are still intact. This procedure is used in vaccinating against tetanus, botulism, and other similar toxic diseases. And, finally, a person can be vaccinated by infecting him with live organisms that have been "attenuated." That is, these organisms either have been grown in special culture mediums or have been passed through a series of animals until they have mutated enough that they will not cause disease but do still carry the specific antigens. This procedure is used to protect against poliomyelitis, yellow fever, measles, smallpox, and many other viral diseases.

PASSIVE IMMUNITY

Thus far, all the acquired immunity that we have discussed has been *active immunity*. That is, the person's body develops either *antibodies* or sensitized lymphocytes in response to invasion of the body by a foreign antigen However, temporary immunity can be achieved in a person without injecting any antigen whatsoever. This is done by infusing antibodies, activated T cells, or both from someone else or from some other animal that has been actively immunized against the antigen. The antibodies will last for two to three weeks, and during that time the person is protected against the invading disease. Activated T cells will last for a few weeks if transfused from another person, and for a few hours to a few days if transfused from an animal. The transfusion of antibodies or lymphocytes to confer immunity is called *passive immunity*.

ALLERGY

One of the important side effects of immunity is development, under some conditions, of allergy. There are several different types of allergy, some of wh

occur in any person, and others that occur only in persons who have a specific allergic tendency.

AN ALLERGY THAT OCCURS IN NORMAL PEOPLE: DELAYED-REACTION ALLERGY

This type of allergy frequently causes skin eruptions in response to certain drugs or chemicals, particularly some cosmetics and household chemicals, to which one's skin is often exposed. Another example of such an allergy is the skin eruption caused by exposure to poison ivy.

Delayed-reaction allergy is caused by activated T cells and not by antibodies. In the case of poison ivy, the toxin of poison ivy in itself does not cause much harm to the tissues. However, upon repeated exposure it does cause the formation of activated T cells. Then, following subsequent exposure to the poison ivy toxin, within a day or so the activated T cells diffuse in sufficient numbers into the skin to combine with the poison ivy toxin and elicit a cell-mediated type of immune reaction. Remembering that this type of immunity can cause release of many toxic substances from the activated T cells as well as extensive invasion of the tissues by macrophages and their subsequent effects, one can well understand that the eventual result of some delayed-reaction allergies can be serious tissue damage.

ALLERGIES IN THE "ALLERGIC" PERSON

Some persons have an "allergic" tendency. Their allergies are called *atopic allergies* because they are caused by a nonordinary response of the immune system. The allergic tendency is genetically passed on from parent to child, and it is characterized by the presence of large quantities of IgE *antibodies*. These antibodies are called *reagins* or *sensitizing antibodies* to distinguish them from the more common IgG antibodies. When an *allergen* (defined as an antigen that reacts specifically with a specific type of IgE reagin antibody) enters the body, an allergen-reagin reaction takes place, and a subsequent allergic reaction takes place.

A special characteristic of the IgE antibodies (the reagins) is a strong propensity to attach to mast cells and basophils. Indeed, these cells can bind as many as half a million molecules of IgE antibodies. Then, when an antigen (an allergen) that has multiple binding sites binds with several of the IgE antibodies attached to a single mast cell or basophil, this causes an immediate change in the membrane of the cell, perhaps resulting from a simple physical effect of the antibody molecules being pulled together by the antigen. At any rate, many of the mast cells and basophils rupture, others release their granules without rupturing and also secrete additional substances not already preformed in the granules. Some of the many different substances that are either released immediately or secreted shortly thereafter include *histamine*, *slow-reacting substance of anaphylaxis*, *eosinophil chemotactic substance*, a *protease*, a *neutrophil chemotactic substance*, *heparin*, and *platelet activating factors*. These substances cause such phenomena as dilatation of the local blood vessels, attraction of eosinophils and neutrophils to the reactive site, damage to the local tissues by the protease, increased permeability of the capillaries and loss of fluid into the tissues, and contraction of local smooth muscle cells. Therefore, any number of different types of abnormal tissue responses can occur, depending upon the type of tissue in which the allergen-reagin reaction occurs. Among the different types of allergic reactions caused in this manner are:

Anaphylaxis. When a specific allergen is injected directly into the circulation it can react in widespread areas of the body with the basophils of the blood and the mast cells located immediately outside the small blood vessels. Therefore, a widespread allergic reaction occurs throughout the vascular system and in closely associated tissues. This is called *anaphylaxis*. The *histamine* released into the circulation causes widespread peripheral vasodilatation as well as increased permeability of the capillaries and marked loss of plasma from the circulation. Often, persons experiencing this reaction die of circulatory shock within a few minutes unless treated with norepinephrine to oppose the effects of the histamine. But also released from the cells is the substance called *slow-reacting substance of anaphylaxis*, which sometimes causes spasm of the smooth muscle of the bronchioles, eliciting an asthma-like attack and sometimes causing death by suffocation.

Urticaria. Urticaria results from antigen entering specific skin areas and causing localized anaphylactoid reactions. *Histamine* released locally causes (a) vasodilatation that induces an immediate *red flare* and (b) increased permeability of the capillaries that leads to swelling of the skin in another few minutes. The swellings are commonly called "hives." Administration of antihistamine drugs to a person prior to exposure will prevent the hives.

Hay Fever. In hay fever, the allergen-reagin reaction occurs in the nose. *Histamine* released in response to this causes local vascular dilatation with resultant increased capillary pressure, and it also causes increased capillary permeability. Both of these effects cause rapid fluid leakage into the tissues of the nose, and the nasal linings become swollen and secretory. Here again, use of antihistamine drugs can prevent this swelling reaction. However, other products of the allergen-reagin reaction still cause irritation of the nose, still eliciting the typical sneezing syndrome despite drug therapy.

Asthma. In asthma, the allergen-reagin reaction occurs in the bronchioles of the lungs. Here, the most important product released from the mast cells seems to be the slow-reacting substance of anaphylaxis, which causes spasm of the bronchiolar smooth muscle. Consequently, the person has difficulty breathing until the reactive products of the allergic reaction have been removed. Unfortunately, administration of antihistamines has little effect on the course of asthma, because histamine does not appear to be the major factor eliciting the asthmatic reaction.

REFERENCES

Amos, D. B., *et al.* (eds.): Immune Mechanisms and Disease. New York, Academic Press, 1979.

Atassi, M. Z.: Antigenic structures of proteins. *Eur. J. Biochem.*, *145*:1, 1984.

Baram, P., *et al.* (eds.): Immunologic Tolerance and Macrophage Function. New York, Elsevier North Holland, 1979.

Barrett, J. T.: Textbook of Immunology: An Introduction to Immunochemistry and Immunobiology. St. Louis: The C. V. Mosby Co., 1983.

Beer, D. J., *et al.*: The influence of histamine on immune and inflammatory responses. *Adv. Immunol.*, *35*:209, 1984.

Benacerraf, B., and Unanue, E. R.: Textbook of Immunology. Baltimore, Williams & Wilkins, 1979.

Bloch, E. F., and Malveaux, F. J.: The significance of immunoglobulin E in resistance to parasitic infections. *Ann. Allergy*, *54*:83, 1985.

Brunt, C, *et al.* (eds.). Systems Theory in Immunology New York, Springer-Verlag, 1979
Caterson, B, *et al*: Production and characterization of monoclonal antibodies directed against connective tissue proteoglycans *Fed Proc*, *44* 386, 1985
Cohen, A S. (ed) Rheumatology and Immunology New York, Grune & Stratton, 1979.
Cooper, J, *et al*: Network regulation among T cells Qualitative and quantitative studies on suppression in the non-immune state *Immunol Rev*, *79* 63, 1984
Corbel, C, and Melebers, F. The synergism of accessory cells and of soluble alpha factors derived from them in the activation of B cells to proliferation. *Immunol Rev*, *78* 51, 1984
Coutinho, A, *et al* T-cell-dependent B cell activation. *Immunol Rev*, *78* 211, 1984
DeLisi, C: Mathematical modeling in immunology *Annu Rev Biophys Bioeng*, *12* 117, 1983
Doherty, P. C, *et al*: Immunological surveillance of tumors in the context of major histocompatibility complex restriction of T cell formation *Adv Cancer Res*, *42* 1, 1984
Ford Hutchinson, A. W Leukotrienes Their formation and role as inflammatory mediators *Fed Proc*, *44* 25, 1985
Forman, J T cells, the MHC, and function *Immunol Rev*, *81* 203, 1984
Frosse, A., Receptors for IgE on mast cells and basophils *Prog Allergy*, *34* 142, 1984
Fudenberg, H H, and Smith, C. L (eds) The Lymphocyte in Health and Disease New York, Grune & Stratton, 1979
Fudenberg, H H, and Whitten, H D Immunostimulation Synthetic and biological modulators of immunity *Annu Rev Pharmacol Toxicol*, *24* 147, 1984
Gatenby, P A, *et al* T cells, T cell subsets and immunoregulation *Aust N.Z J Med*, *14* 89, 1984
Gerrard, J W Food Allergy Springfield, Ill, Charles C Thomas, 1980
Glick, B. Interrelation of the avian immune and neuroendocrine systems *J Exp Zool*, *232* 671, 1984
Hern, A. W, *et al*: Blood Cell Formation and the Cellular Basis of Immune Responses Philadelphia, J.B Lippincott Co, 1979
Hobbs, J R (ed.) Thymic Factor Therapy New York, Raven Press, 1985.
Horsberger, M A, and Hochkeppel, H K An interferon induced mouse protein involved in the mechanism of resistance to influenza viruses Its purification to homogeneity and characterization by polyclonal antibodies *J. Biol. Chem*, *260* 1730, 1985
Johnson, F. (ed) Allergy, Including IGE in Diagnosis and Treatment Chicago, Year Book Medical Publishers, 1979
Kaplan, J G (ed) The Molecular Basis of Immune Cell Function New York, Elsevier/North-Holland, 1979
Kay, M. M: Immunological aspects of aging Early changes in thymic activity *Mech Ageing Dev*, *30* 193, 1984
Kendall, M D *Have we underestimated the importance of the thymus in man? Experientia*, *40* 1181, 1984
Kimball, J W Introduction to Immunology New York, Macmillan Publishing Co, Inc, 1983.
Koffler, D: The Immunology of Rheumatoid Diseases Summit, N.J., CIBA Pharmaceutical Company, 1979
Kouttab, N. M, *et al*: Lymphokines and monokines as regulators of human lymphoproliferation *Clin Chem*, *30* 1539, 1984
Lancki, D W, *et al* Cell surface structures involved in T cell activation *Immunol Rev*, *81* 65, 1984
Langer, J A, and Pestka, S Purification, bacterial expression, and biological activities of the human interferons *J. Invest Dermatol*, *83* (Suppl 1) 128s, 1984
Levy, R, and Miller, R A. Biological and clinical implications of lymphocyte hybridomas Tumor therapy and monoclonal antibodies *Annu Rev Med*, *34* 107, 1983
Litman, G. W., and Good, R A (eds.) Immunoglobulins New York, Plenum Press, 1977
Malmsten, C L Leukotrienes Mediators of inflammation and immediate hypersensitivity reactions *CRC Crit Rev Immunol*, *4* 307, 1984
Marchesi, V T, *et al* (eds) Cell Surface Carbohydrates and Biological Recognition New York, A R. Liss, 1978
Martinez, C, *et al*: Induction of B lymphocytes by helper T cells Requirement of direct cell contact. *Ann Immunol. (Paris)*, *135D* 82, 1984
McGiven, A. R. (ed) Immunological Investigation of Renal Disease New York, Churchill Livingstone, 1980
McNamara, M, *et al* T-cell helper circuits *Immunol Rev*, *79* 87, 1984
Metzger, H: Early molecular events in antigen-antibody cell activation *Annu Rev Pharmacol Toxicol*, *19* 427, 1979
Murayama, A, *et al*: Regulation of human B-cell activation, proliferation, and differentiation by soluble factors *J Clin Immunol*, *4* 337, 1984
Nahmias, A J, and O'Reilly, R (eds.) Immunology of Human Infection New York, Plenum Press, 1979
Neuwenhuis, P (ed) In Vivo Immunology Histophysiology of the Lymphoid System. New York, Plenum Publishing Corp, 1982
Norcross, M A. A synaptic basis for T-lymphocyte activation *Ann Immunol (Paris)*, *135D* 113, 1984
Paraskevas, F Pathways of T cell suppression *CRC Crit Rev Immunol*, *5* 95, 1984
Perrin, B, and Vogel, H J (eds.) Cells of Immunoglobulin Synthesis. New York, Academic Press, 1979
Pinchera, A, and Fenzi, G Endocrine autoimmune diseases *In* De-Groot, L J, *et al* (eds) Endocrinology Vol 3 New York, Grune & Stratton, 1979, p 2063
Piper, P J Formation and actions of leukotrienes *Physiol Rev*, *64* 744, 1984
Pollard, B, *et al.* (eds.) Inborn Errors of Specific Immunity New York, Academic Press, 1979
Proffitt, M R (ed) Virus-Lymphocyte Interactions Implications of Disease New York, Elsevier-North Holland, 1979
Salton, M R J (ed) Immunochemistry of Enzymes and Their Antibodies Huntington, N Y, R E Krieger Publishing Co, 1980.
Samter, M (ed) Immunological Diseases Boston, Little, Brown, 1978
Schiff, G M Active immunization for adults *Annu Rev Med*, *31* 441, 1980
Schwartz, L M Compendium of Immunology New York, Van Nostrand Reinhold, 1979
Schwartz, L M (ed) Compendium of Immunology New York, Van Nostrand Reinhold, 1983
Sercarz, E E, and Cunningham, A J (eds) Strategies of Immune Regulation New York, Academic Press, 1979
Shannon, K M., and Ammann, A J Acquired immune deficiency syndrome in childhood *J Pediatr 106* 332 1985
Simonsen, M, *et al* Compound receptors in the cell membrane Ruminations from the borderland of immunology and physiology *Prog Allergy*, *36* 151, 1985
Smith, W L, and Garcia-Perez A Immunodissection Use of monoclonal antibodies to isolate specific types of renal cells *Am J Physiol*, *248* F1, 1985
Stadler, B M, and de Weck, A L Role of lymphokines in immediate-type allergy *Springer Semin Immunopathol*, *7* 415, 1984
Stuart, F P, and Fitch, F W, (eds) Immunological Tolerance and Enhancement Baltimore, University Park Press, 1979
Takemore, T, and Rajewsky, K Mechanism of neonatally induced idiotype suppression and its relevance for the acquisition of self-tolerance *Immunol Rev*, *79* 103, 1984
Taylor, D W, and Siddiqui, W A Recent advances in malarial immunity *Annu Rev Med*, *33* 69, 1982
Terry, W D, and Yamamura, Y (eds.) Immunobiology and Immunotherapy of Cancer New York, Elsevier/North-Holland, 1979.
Vitetta, E S, *et al* T-cell-derived lymphokines that induce IgM and IgG secretion in activated murine B cells *Immunol. Rev*, *78* 137, 1984
Wybran, J Enkephalins and endorphins as modifiers of the immune system Present and future *Fed Proc*, *44* 92, 1985

7

Blood Groups; Transfusion; Tissue and Organ Transplantation

ANTIGENICITY AND IMMUNE REACTIONS OF BLOOD

When blood transfusions from one person to another were first attempted the transfusions were successful in some instances, but in many more immediate or delayed agglutination and hemolysis of the red blood cells occurred. Soon it was discovered that the bloods of different persons usually have different antigenic and immune properties, so that antibodies in the plasma of one blood react with antigens on the surfaces of the red cells of another Furthermore, the antigens and the antibodies are almost never precisely the same in one person as in another For this reason, it is easy for blood from a donor to be mismatched with that of a recipient. Fortunately, if proper precautions are taken, one can determine ahead of time whether or not appropriate antibodies and antigens are present in the donor and recipient bloods to cause a reaction, but, on the other hand, lack of proper precautions often results in varying degrees of red cell agglutination and hemolysis, resulting in a typical transfusion reaction that can lead to death.

Multiplicity of Antigens in the Blood Cells. At least 30 commonly occurring antigens, each of which can at times cause antigen-antibody reactions, have been found in human blood cells, especially on the surfaces of the cell membranes In addition to these, more than 300 others of less potency or that occur in individual families rather than having widespread occurrence are known to exist Among the 30 or more common antigens, certain ones are highly antigenic and regularly cause transfusion reactions if proper precautions are not taken, whereas others are of importance principally for studying the inheritance of genes and therefore for establishing parentage, race, and so forth Essentially all these antigens are either glycolipids or glycoproteins

Bloods are divided into different *groups* and *types* in accordance with the types of antigens present in the cells Two particular groups of antigens are more likely than the others to cause blood transfusion reactions These are the so-called *O-A-B* system of antigens and the *Rh* system

O-A-B BLOOD GROUPS

THE A AND B ANTIGENS—CALLED "AGGLUTINOGENS"

Two related antigens—type A and type B—occur on the surfaces of the red blood cells in a large proportion of the population Because of the way these antigens are inherited, people may have neither of them on their cells, they may have one, or they may have both simultaneously

As will be discussed subsequently, strong antibodies that react specifically with either the type A or type B antigen almost always occur in the plasmas of persons who do not have the antigens on their red cells. These antibodies bind with the red cell antigens to cause agglutination of the red cells Therefore, the type A and type B antigens are called *agglutinogens,* and the plasma antibodies that cause the agglutination are called *agglutinins* It is on the basis of the presence or absence of the agglutinogens in the red blood cells that blood is grouped for the purpose of transfusion.

The Four Major O-A-B Blood Groups. In transfusing blood from one person to another, the bloods of donors and recipients are normally classified into four major O-A-B groups, as illustrated in Table 7–1, depending on the presence or absence of the two agglutinogens When neither A nor B agglutinogen is present the blood group is *group O* When only type A agglutinogen is present, the blood is *group A* When only type B agglutinogen is present, the blood is *group B* And when both A and B agglutinogens are present, the blood is *group AB*

Relative Frequencies of the Different Blood Types. The prevalence of the different blood types among caucasoids is approximately as follows

Type	Per cent
O	47
A	41
B	9
AB	3

It is obvious from these percentages that the O and A genes occur frequently whereas the B gene is infrequent

Genetic Determination of the Agglutinogens. Genes on two adjacent chromosomes, one gene on each chromosome, determine the O-A-B blood groups. These are allelomorphic genes that can be any one of three different types, but only one type on each chromosome: type O, type A, or type B. There is no dominance among the three different allelomorphs. However, the type O gene is either functionless or almost functionless, so that it causes either no type O agglutinogen in the cells or such a weak agglutinogen that it is normally insignificant. On the other hand, the type A and type B genes do cause strong agglutinogens in the cells. Therefore, if either of the genes on the two respective chromosomes is type A, the red blood cells will contain type A *agglutinogen*, and likewise, if either of the two genes is type B, the red blood cells will contain type B agglutinogen. Or, if the gene on one chromosome is type A and on the other is type B, the red blood cells will contain both A and B agglutinogens.

The six possible combinations of genes, as shown in Table 7–1, are OO, OA, OB, AA, BB, and AB. These different combinations of genes are known as the *genotypes*, and each person is one of the six different genotypes.

One can observe from tha table that a person with genotype OO produces no agglutinogens at all, and, therefore, the blood group is O. A person with either genotype OA or AA produces type A agglutinogens and, therefore, has blood group A. Genotypes OB and BB give group B blood, and genotype AB gives group AB blood.

THE AGGLUTININS

When type A agglutinogen *is not present* in a person's red blood cells, antibodies known as "anti-A" agglutinins develop in his plasma. Also, when type B agglutinogen *is not present* in the red blood cells, antibodies known as "anti-B" agglutinins develop in the plasma.

Thus referring once again to Table 7–1, it will be observed that group O blood, though containing no agglutinogens, does contain both *anti-A* and *anti-B* agglutinins, while group A blood contains type A agglutinogens and *anti-B agglutinins*, and group B blood contains type B agglutinogens and *anti-A agglutinins*. Finally, group AB blood contains both A and B agglutinogens but no agglutinins at all.

Titer of the Agglutinins at Different Ages. Immediately after birth the quantity of agglutinins in the plasma is almost zero. Two to eight months after birth, the infant begins to produce agglutinins—anti-A agglutinins when type A agglutinogens are not present in the cells and anti-B agglutinins when type B agglutinogens are not in the cells. Figure 7–1 illustrates the changing

Table 7–1 THE BLOOD GROUPS WITH THEIR GENOTYPES AND THEIR CONSTITUENT AGGLUTINOGENS AND AGGLUTININS

Genotypes	Blood Groups	Agglutinogens	Agglutinins
OO	O	—	Anti-A and Anti-B
OA or AA	A	A	Anti-B
OB or BB	B	B	Anti-A
AB	AB	A and B	—

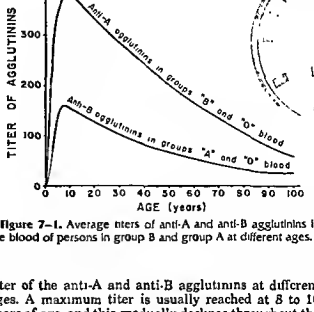


Figure 7–1. Average titers of anti-A and anti-B agglutinins in the blood of persons in group B and group A at different ages.

titer of the anti-A and anti-B agglutinins at different ages. A maximum titer is usually reached at 8 to 10 years of age, and this gradually declines throughout the remaining years of life.

Origin of Agglutinins in the Plasma. The agglutinins are gamma globulins, as are other antibodies, and are produced by the same cells that produce antibodies to any other antigens. Most of them are IgM and IgG immunoglobulin molecules.

But why are these agglutinins produced in individuals who do not have the antigenic substances in their red blood cells? The answer to this seems to be that small amounts of group A and B antigens enter tha body in the food, in bacteria, and in other ways, and these substances initiate the development of the anti-A or anti-B agglutinins. One of the reasons for believing this is that injection of group A or group B antigen into a recipient having another blood type causes a typical immune response with formation of greater quantities of agglutinins than ever. Also, the newborn baby has few if any agglutinins, showing that agglutinin formation occurs almost entirely after birth.

THE AGGLUTINATION PROCESS IN TRANSFUSION REACTIONS

When bloods are mismatched so that anti-A or anti-B agglutinins are mixed with red blood cells containing A or B agglutinogens respectively, the red cells agglutinate by the following process: The agglutinins attach themselves to the red blood cells. Because the agglutinins have two binding sites (IgG type) or ten sites (IgM type), a single agglutinin can attach to two or more different red blood cells at the same time, thereby causing the cells to adhere to each other. This causes the cells to clump. Then these clumps plug small blood vessels throughout the circulatory system. During the ensuing few hours to few days, the phagocytic white blood cells and the reticuloendothelial system destroy the agglutinated cells, releasing hemoglobin into the plasma.

Hemolysis in Transfusion Reactions. Sometimes, when recipient and donor bloods are mismatched, immediate hemolysis of red cells occurs in the circulating blood. In this case the antibodies cause lysis of the red blood cells by activating the complement system. This

in turn releases proteolytic enzymes (the *lytic complex*) that rupture the rell membranes, as was described in Chapter 6.

However, immediate intravascular hemolysis is far less common than agglutination, because not only does there have to be a very high titer of antibodies for this to occur but also a different type of antibody seems to be required, mainly the IgM antibodies, these antibodies are called *hemolysins* However, even agglutination eventually leads to hemolysis of the agglutinated red cells because the phagocytic white blood cells and the reticuloendothelial cells rupture the agglutinated cells within a few hours after agglutination has taken place

BLOOD TYPING

Prior to giving a transfusion, it is necessary to determine the blood group of the recipient and the group of the donor blood so that the bloods will be appropriately matched. This is called *blood typing*, and it is performed in the following way The red blood cells are first diluted with saline. One portion is then mixed with anti-A agglutinin while another portion is mixed with anti-B agglutinin After several minutes, the mixture is observed under a microscope. If the red blood cells have become clumped—that is, "agglutinated"—one knows that an antibody-antigen reaction has resulted

Table 7-2 illustrates the reactions that occur with each of the four different types of blood. Group O red blood cells have no agglutinogens and, therefore, do not react with either the anti-A or the anti-B serum. Group A blood has A agglutinogens and therefore agglutinates with anti-A agglutinins Group B blood has B agglutinogens and agglutinates with anti-B serum. Group AB blood has both A and B agglutinogens and agglutinates with both types of serum.

THE Rh BLOOD TYPES

In addition to the O-A-B blood group system, several other systems are sometimes important in the transfusion of blood The most important of these is the Rh system. The one major differenre between the O-A-B system and the Rh system is the following: In the O-A-B system, the agglutinins responsible for causing transfusion reactions develop spontaneously, while in the Rh system spontaneous agglutinins almost never occur Instead, the person must first be massively exposed to an Rh antigen, usually by transfusion of blood, before enough agglutinins to cause a significant transfusion reaction will develop

The Rh Antigens—"Rh Positive" and "Rh Negative" Persons. There are six common types of Rh antigens, each of which is called an Rh factor. These types are designated C, D, E, c, d, and e. A person who has a C antigen will not have the c antigen, but the person missing the C antigen will always have the c antigen. The same is true for the D-d and E-e antigens Also, because of the manner of inheritance of these factors, each person will have one of each of the three pairs of antigens.

The type D antigen is widely prevalent in the population, and it is also considerably more antigenic than the other Rh antigens Therefore, anyone who has this type of antigen is said to be *Rh positive*, while those persons who do not have type D antigen are said to be *Rh negative*. However, it must be noted that even in Rh negative persons some of the other Rh antigens can still cause transfusion reactions.

Approximately 85 per cent of all caucasoids are Rh positive and 15 per cent, Rh negative In American Blacks, the percentage of Rh positives is about 95, whereas in African Blacks it is virtually 100 per cent.

Typing Bloods for Rh Factors. Typing for Rh factors is performed in a manner similar to that used in typing the A-B-O agglutinogens. This is usually achieved using from 4 to 6 different anti-Rh sera However, the anti-Rh antibodies are far less potent in their capability of causing red cell agglutination than are the anti-A and the anti-B antibodies Therefore, to cause agglutination in the presence of the Rh antibodies, a small amount of protein must be added to the reactant mixture; this protein provides cross-linkages between the antibodies after they attoch to the red blood cells.

Table 7-2 BLOOD TYPING—SHOWING AGGLUTINATION OF CELLS OF THE DIFFERENT BLOOD GROUPS WITH ANTI-A AND ANTI-B AGGLUTININS

Red Blood Cells	Sera	
	Anti-A	*Anti B*
O	–	–
A	+	–
B	–	+
AB	+	+

THE Rh IMMUNE RESPONSE

Formation of Anti-Rh Agglutinins. When red blood cells containing Rh factor, or even protein breakdown products of such cells, are injected into a person without the factor—that is, into the Rh negative person—anti-Rh agglutinins develop very slowly, the maximum concentration of agglutinins occurring approximately two to four months later. This immune response occurs to a much greater extent in some people than in others. On multiple exposure to the Rh factor, the Rh negative person eventually becomes strongly "sensitized" to the Rh factor—that is, he or she develops a very high titer of anti-Rh agglutinins

Characteristics of Rh Transfusion Reactions. If an Rh negative person has never before been exposed to Rh positive blood, transfusion of Rh positive blood into him causes no immediate reaction at all However, in some of these persons anti-Rh antibodies develop in sufficient quantities during the next two to four weeks to cause agglutination of the transfused cells that are still in the blood These cells are then hemolyzed by the tissue macrophage system. Thus, a delayed transfusion reaction occurs, though it is usually mild. Yet, on subsequent transfusion of Rh positive blood into the same person, who is now immunized against the Rh factor, the transfusion reaction is greatly enhanced and can be as severe as the reactions that occur with types A and B bloods

Erythroblastosis Fetalis. Erythroblastosis fetalis is a disease of the fetus and newborn infant characterized by progressive agglutination and subsequent phagocytosis of the red blood cells In most instances of erythroblastosis fetalis the mother is Rh negative, and the father is Rh positive; the baby has inherited the Rh positive characteristic from the father, and the mother

has developed anti-Rh agglutinins that have diffused through the placenta into the fetus to cause red blood cell agglutination.

Prevalence of the Disease. An Rh negative mother having her first Rh positive child usually does not develop sufficient anti-Rh agglutinins to cause any harm. However, an Rh negative mother having her second Rh positive child often will have become "sensitized" by the first child and therefore will often develop anti-Rh agglutinins rapidly upon becoming pregnant with the second child. Approximately 3 per cent of these second babies exhibit some signs of erythroblastosis fetalis; approximately 10 per cent of the third babies exhibit the disease; and the incidence rises progressively with subsequent pregnancies.

The Rh negative mother develops anti-Rh agglutinins only when the fetus is Rh positive. Many of the Rh positive fathers are heterozygous (about 55 per cent), causing about one fourth of the offspring to be Rh negative. Therefore, after an erythroblastotic child has been born, it is not certain that future children will also be erythroblastotic.

Effect of the Mother's Antibodies on the Fetus. After anti-Rh antibodies have formed in the mother, they diffuse very slowly through the placental membrane into the fetus' blood. There they cause slow agglutination of the fetus' blood. The agglutinated red blood cells gradually hemolyze, releasing hemoglobin into the blood. The macrophages then convert the hemoglobin into bilirubin, which causes yellowness (jaundice) of the skin. The antibodies probably also attack and damage many of the other cells of the body.

Clinical Picture of Erythroblastosis. The newborn, jaundiced, erythroblastotic baby is usually anemic at birth, and the anti-Rh agglutinins from the mother usually circulate in the baby's blood for one to two months after birth, destroying more and more red blood cells.

The hemopoietic tissues of the baby attempt to replace the hemolyzing red blood cells. The liver and the spleen become greatly enlarged and produce red blood cells in the same manner that they normally do during the middle of gestation. Because of the very rapid production of cells, many early forms, including many nucleated blastic forms, are emptied into the circulatory system, and it is because of the presence of these in the blood that the disease has been called "erythroblastosis fetalis."

Though the severe anemia of erythroblastosis fetalis is usually the cause of death, many children who barely survive from the anemia exhibit permanent mental impairment or damage to motor areas of the brain because of precipitation of bilirubin in the neuronal cells, causing their destruction, a condition called *kernicterus.*

Treatment of the Erythroblastotic Baby. The usual treatment for erythroblastosis fetalis is to replace the newborn infant's blood with Rh negative blood. Approximately 400 ml of Rh negative blood is infused over a period of 1.5 or more hours while the baby's own Rh positive blood is being removed. This procedure may be repeated several times during the first few weeks of life, mainly to keep the bilirubin level low and thereby to prevent kernicterus. By the time these transfused Rh negative cells are replaced with the baby's own Rh positive cells, a process that requires six or more weeks, the anti-Rh agglutinins that had come from the mother will have been destroyed.

OTHER BLOOD FACTORS

Many antigenic proteins besides the O, A, B, and Rh factors are present in red blood cells of different persons, but these other factors only rarely cause transfusion reactions and, therefore, are mainly of academic and legal importance. Some of these different blood factors are the M, N, S, s, P, Kell, Lewis, Duffy, Kidd, Diego, and Lutheran factors. Occasionally, multiple transfusions of red cells containing one of these factors cause specific agglutinins to develop in the recipient, and transfusion reactions can then occur when the same type of blood is infused again.

Method for Studying Obscure Blood Factors. One of the means by which different blood factors, including the Rh factor, have been studied in the human being has been to immunize lower animals, such as rabbits, with human blood cells. Specific antibodies develop in the animal against the antigens of the human cells, and specific immune sera prepared from the plasma of these animals can then be used for determining the presence or absence of the same antigens in red blood cells of other persons

Blood Typing in Legal Medicine. In the past three decades the use of blood typing has become an important legal procedure in cases of disputed parentage. Including all the blood types, as many as 50 common blood group genes can be determined for each person by blood-typing procedures. After the mother's and child's genes have been determined, many of the father's genes and his corresponding blood factors are then known immediately, because any gene present in a child but not present in the mother must be present in the father. If a suspected man is missing any one of the necessary blood factors, he is not the father. A falsely charged man can be cleared in about 75 per cent of the cases using commonly available antisera for blood typing, and in essentially all cases using all possible types of antisera.

TRANSFUSION

Indications for Transfusion. The most common reason for transfusion is decreased blood volume, which will be discussed in detail in Chapter 28 in relation to shock. Also, transfusions are often used for treating anemia or to supply the recipient with some other constituent of whole blood besides red blood cells, such as to supply a thrombocytopenic patient with new platelets. Also, hemophilic patients can be rendered temporarily nonhemophilic by plasma transfusion, and, occasionally, the quantity of "complement" in a recipient must be supplemented by fresh plasma infusions before certain antigen-antibody reactions can take place.

TRANSFUSION REACTIONS RESULTING FROM MISMATCHED BLOOD GROUPS

If blood of one blood group is transfused to a recipient of another blood group, a transfusion reaction is likely to occur in which the red blood cells *of the donor blood* are agglutinated. It is very rare that the transfused blood ever causes agglutination *of the recipient's cells* for the following reason: The plasma portion of the donor blood immediately becomes diluted by all the plasma of the recipient, thereby decreasing the titer of the

agglutinins to a level too low to cause agglutination. On the other hand, the infused blood does not dilute the agglutinins in the recipient's plasma to a major extent. Therefore, the recipient's agglutinins can still agglutinate the donor cells

Hemolysis of Red Cells Following Transfusion Reactions. All transfusion reactions resulting from mismatched blood groups eventually cause hemolysis of the red blood cells. Occasionally, the antibodies are potent enough and are composed of the appropriate class of immunoglobulins to cause immediate hemolysis, but more frequently the cells agglutinate first and then are mainly entrapped in the peripheral vessels. Over a period of hours to days the entrapped cells are phagocytized, thereby liberating hemoglobin into the circulatory system

When the rate of hemolysis is rapid, the concentration of hemoglobin in the plasma can rise to extremely high values A small quantity of hemoglobin can become attached to one of the plasma proteins, *haptoglobin*, and continue to circulate in the blood without causing any harm. However, above the threshold value of 100 mg of hemoglobin per 100 ml of plasma, the excess hemoglobin remains in the free form and diffuses out of the circulation into the tissue spaces or through the renal glomeruli into the kidney tubules, as is discussed below. The hemoglobin remaining in the circulation or passing into the tissue spaces is gradually ingested by phagocytic cells and converted into bilirubin and later excreted into the bile by the liver, which will be discussed in Chapter 70. The concentration of bilirubin in the body fluids often rises high enough to cause *jaundice*—that is, the person's tissues become tinged with yellow pigment. But, if liver function is normal, jaundice usually does not appear unless more than 300 to 500 ml of blood is hemolyzed in less than a day.

Acute Kidney Shutdown Following Transfusion Reactions. One of the most lethal effects of transfusion reactions is acute kidney shutdown, which can begin within a few minutes to a few hours and continue until the person dies of renal failure.

The kidney shutdown seems to result from three different causes First, the antigen-antibody reaction of the transfusion reaction releases toxic substances from the hemolyzing blood that cause powerful renal vasoconstriction Second, the loss of circulating red cells along with production of toxic substances from the hemolyzed cells and from the immune reaction often causes circulatory shock, the arterial blood pressure falls very low and the renal blood flow and urinary output decrease Third, if the total amount of free hemoglobin in the circulating blood is greater than that quantity which can bind with haptoglobin, much of the excess leaks through the glomerular membranes into the kidney tubules If this amount is still slight, it can be reabsorbed through the tubular epithelium into the blood and will cause no harm, but, if it is great, then only a small percentage is reabsorbed Yet water continues to be reabsorbed, causing the tubular hemoglobin concentration to rise so high that it precipitates and blocks many of the tubules, this is especially true if the urine is acidic Thus, renal vasoconstriction, circulatory shock, and tubular blockage all add together to cause acute renal shutdown If the shutdown is complete, the patient dies within a week to 12 days, as explained in Chapter 38, unless treated with the artificial kidney.

TRANSPLANTATION OF TISSUES AND ORGANS

Relation of Genotypes to Transplantation. In thi modern age of surgery, many attempts are being mad to transplant tissues and organs from one person t another, or, occasionally, from lower animals to th human being. *Many of the different antigenic protein* of red blood cells that cause transfusion reactions plu still many more are present in the other cells of th body as well. Consequently, any foreign cells trans planted into a recipient can cause immune response and immune reactions. In other words, most recipient are just as able to resist invasion by foreign cells as t resist invasion by foreign bacteria

Isografts, Allografts, and Xenografts. A transplan of a tissue or whole organ from one identical twin t another is called an *isograft*. A transplant from on human being to another or from any animal to anothe animal of the same species is called an *allograft* Finally a transplant from a lower animal to a human being o from an animal of one species to one of another specie is called a *xenograft*.

Transplantation of Cellular Tissues. In the case c isografts, cells in the transplant will almost always liv indefinitely if an adequate blood supply is provided, bu in the case of allografts and xenografts, immune reac tions almost always occur, causing death of all the cell in the graft within two to five weeks after transplanta tion unless some specific therapy is used to prevent th immune reaction The cells of allografts usually persis much longer than those of xenografts because the anti genic makeup of the allograft is more nearly the sam as that of the recipient's tissues than is true for th xenograft. And when the tissues are properly "typed and are very similar prior to transplant, completel successful allografts occasionally result

Some of the different cellular tissues and organs tha have been transplanted either experimentally or fo therapeutic benefit from one person to another are skin kidney, heart, liver, glandular tissue, bone marrow, an lung. Many kidney allografts have been successful fo as long as five to ten years, and a few liver and hear transplants for one to ten years

Transplantation of Noncellular Tissues. Certai tissues that have no cells or in which the cells ar unimportant to the purpose of the graft, such as th cornea, tendon, fascia, and bone, can often be graftec from one person to another with considerable success In these instances the grafts act merely as a supportin latticework into which or around which the surroundin living tissues of the recipient grow Indeed, some suc grafts—bone and fascia—are occasionally successfu even when they come from a lower animal rather tha from another human being.

ATTEMPTS TO OVERCOME THE ANTIGEN-ANTIBODY REACTIONS IN TRANSPLANTED TISSUE

Because of the extreme potential importance of transplanting certain tissues and organs, such as skin, kidneys, and lungs, serious attempts have been made to prevent the antigen-antibody reactions associated with

transplants. The following specific procedures have met with certain degrees of clinical or experimental success.

Tissue Typing. In the same way that red blood cells can be typed to prevent reactions between recipient and donor, so also is it possible to "type" tissues to help prevent graft rejection, though thus far this procedure has met with far less success than has been achieved in red blood cell typing.

The most important antigens that cause graft rejection are a group of antigens called the HLA antigens. This group of about 100 different antigens in the tissue cell membranes are determined by four separate genes at the so-called *HLA genetic locus*. The genes are allelomorphic and, therefore, can code for only four of the HLA antigens in any one person. That is, only four of the antigens can occur in any one individual, but even this represents a multitude of possible different combinations.

The HLA antigens occur in the white blood cells as well as in the tissue cells. Therefore, tissue typing for these antigens is done in the membranes of lymphocytes that have been separated from the person's blood. The lymphocytes are mixed with appropriate antisera and complement, and after appropriate incubation the cells are tested for membrane damage, usually by testing the rate of uptake by the lymphocytic cells of a supravital dye.

Fortunately, some of the HLA antigens are not severely antigenic, for which reason precise match of some of the antigens between donor and recipient is not absolutely essential to allow allograft acceptance. Nevertheless, by obtaining the best possible match between donor and recipient, the grafting procedure has become far less hazardous. The best success has been tissue-type matches between members of the same family. Of course, the match in identical twins is exact so that transplants between identical twins almost never reject because of immune reactions.

Prevention of Graft Rejection by Suppressing the Immune System. Obviously, if the immune system were completely suppressed, graft rejection would not occur. In fact, in an occasional person who has serious depression of the immune system, grafts can be successful without the use of any therapy at all to prevent rejection. However, in the normal person, even with the best possible tissue typing, allografts will rarely resist rejection for more than a few weeks to few months without the employment of some therapeutic procedure to suppress the immune system. Furthermore, since the T cells are mainly important for killing the cells of the grafts, their suppression is much more important than suppression of the human antibodies. Some of the therapeutic agents that have been used for this purpose include the following:

(1) *Glucocorticoid hormones (or drugs with glucocorticoid activity)*, that suppress the growth of all lymphoid tissue and therefore decrease the formation of both antibodies and T cells.

(2) *Various drugs that have a toxic effect on the lymphoid system* and therefore block the formation of antibodies and T cells, especially the drug azathioprine (Imuran).

(3) *Cyclosporin A*, that has a specific cytotoxic effect on T cells and therefore is especially efficacious in blocking the rejection reaction.

Unfortunately, use of these agents leaves the person unprotected from disease; therefore, bacterial and viral infections sometimes become rampant. In addition, the incidence of cancer is several times as great in an immunosuppressed person, presumably because the immune system is important in destroying many early cancers before they can really begin to develop.

To summarize, transplantation of living tissues in human beings until the present has been mainly an experiment except in the case of kidney transplants, which are now successful in the majority of the cases for at least a few years, but only when massive immunosuppressive treatment is used simultaneously. On the other hand, when someone succeeds in blocking the immune response of the recipient to a donor organ without at the same time destroying the recipient's specific immunity for disease, the story will change overnight.

REFERENCES

Ascher, N L., *et al* Cellular basis of allograft rejection. *Immunol Rev*, *77* 217, 1984

Ballantyne, D L., and Converse, J M. Experimental Skin Grafts and Transplantation Immunity A Recapitulation. New York, Springer-Verlag, 1979

Chatterjee, B N (ed) Symposium on Organ Transplantation. Philadelphia, W B Saunders Co, 1978

Chatterjee, S N (ed) Renal Transplantation A Multidisciplinary Approach New York, Raven Press, 1980

Claas, F H, and van Rood, J J The interaction of drugs and endogenous substances with HLA class-I antigens. *Prog Allergy*, *36*:135, 1985

Colvin, R B Flow cytometric analysis of T cells. Diagnostic applications in transplantation *Ann N Y Acad Sci*, *428* 5, 1984.

Cunningham, B A. The structure and function of histocompatibility antigens. *Sci Am*, *234*(4) 96, 1977.

Demetris, A J, *et al* Pathology of hepatic transplantation A review of 62 adult allograft recipients immunosuppressed with a cyclosporin/steroid regimen *Am J Pathol* *118* 151, 1985

Doherty, P C, *et al* Immunological surveillance of tumors in the context of major histocompatibility complex restriction of T cell formation *Adv Cancer Res*, *42* 1, 1984

Finch, F W, *et al* Contributions of T cell clones to an understanding of the T cell receptor for antigen *Transplantation*, *38*:201, 1984

Goulmy, E Class-I-restricted human cytotoxic T lymphocytes directed against minor transplantation antigens and their possible role in organ transplantation. *Prog Allergy*, *36* 44, 1985.

Hancock, W W Analysis of intragraft effector mechanisms associated with human renal allograft rejection Immunohistological studies with monoclonal antibodies. *Immunol Rev*, *77* 61, 1984

Hobbs, J R. (ed.) Thymic Factor Therapy New York, Raven Press, 1985

Hobbell, R C (ed) Advances in Blood Transfusion Arlington, Va., American Blood Commission, 1979

Hunt, S A, and Stinson, E B Cardiac transplantation. *Annu. Rev Med*, *32* 213, 1981.

Kahan, B D, *et al*.. Clinical and experimental studies with cyclosporin in renal transplantation. *Surgery*, *97* 125, 1985

Kaufman, J F, *et al* The class II molecules of the human and murine major histocompatibility complex *Cell*, *36* 1, 1984

Mason, D W, *et al*: *Mechanisms of allograft rejection The roles of cytotoxic* T-cells and delayed-type hypersensitivity. *Immunol. Rev.*, *77* 167, 1984

Mohn, J. F., *et al* (eds.) Human Blood Groups New York, S. Karger, 1977

Rosenfeld, R. E., *et al*: Genetic model for the Rh blood group system. *Proc Natl Acad Sci USA*, *70* 1303, 1973

Selwood, N., and Hedges, A. Transplantation Antigens. A Study in Serological Data Analysis. New York, John Wiley & Sons, 1978

Thomson, A. W, *et al*: Cyclosporine Immunology, toxicity and pharmacology in experimental animals *Agents Actions*, *15* 306, 1984.

Touraine, J L (ed) Transplantation and Clinical Immunology New York, Elsevier Science Publishing Co., Inc., 1981

Wood, R. F., Renal Transplantation A Clinical Handbook. Philadelphia, W. B Saunders Co., 1984

Zmijewski, C M Human leukocyte antigen matching in renal transplantation. Review and current status. *J Surg Res*, *35* 56, 1983

8

Hemostasis and Blood Coagulation

EVENTS IN HEMOSTASIS

The term *hemostasis* means prevention of blood loss Whenever a vessel is severed or ruptured, hemostasis is achieved by several different mechanisms, including (1) vascular spasm, (2) formation of a platelet plug, (3) blood coagulation, and (4) eventual growth of fibrous tissue into the blood clot to close the hole in the vessel permanently

VASCULAR SPASM

Immediately after a blood vessel is cut or ruptured, the stimulus of the traumatized vessel causes the wall of the vessel to contract; this instantaneously reduces the flow of blood from the vessel rupture The contraction results from both nervous reflexes and local myogenic spasm The nervous reflexes presumably are initiated by pain or other impulses originating from the traumatized vessel or from nearby tissues. However, most of the spasm probably results from local myogenic contraction of the blood vessels initiated by direct damage to the vascular wall, which presumably causes transmission of action potentials along the vessel wall for several centimeters and results in constriction of the vessel. The more of the vessel that is traumatized, the greater the degree of spasm, this means that a sharply cut blood vessel usually bleeds much more than does a vessel ruptured by crushing. This local vascular spasm can last for many minutes or even hours, during which time the ensuing processes of platelet plugging and blood coagulation can take place

The value of vascular spasm as a mean of hemostasis is illustrated by the fact that persons whose legs have been severed by crushing types of trauma sometimes have such intense spasm in vessels as large as the anterior tibial artery that there is not lethal loss of blood

FORMATION OF THE PLATELET PLUG

If the rent in the blood vessel is very small—and many very small vascular holes do develop each day—it is often sealed by a *platelet plug* rather than by a blood clot. To understand this it is important that we first discuss the nature of platelets themselves.

Physical and Chemical Characteristics of Platelets

Platelets are minute round or oval discs 2 to 4 microns in diameter. They are formed in the bone marrow from *megakaryocytes*, which are extremely large cells of the hemopoietic series in the bone marrow that do not themselves leave the marrow to enter the blood. Instead, platelets form as buds on the surfaces of the megakaryocytes and pinch off to be released into the blood. The normal concentration of platelets in the blood is between 150,000 and 350,000 per cubic millimeter.

Platelets have many functional characteristics of whole cells, even though they do not have nuclei and cannot reproduce In their cytoplasm are such active factors as (1) *actin* and *myosin molecules*, similar to those found in muscle cells, that can cause the platelets to contract, (2) residuals of both the *endoplasmic reticulum* and the *Golgi apparatus* that synthesize various enzymes and store large quantities of calcium ions; (3) enzyme systems that are capable of forming *ATP* and *ADP*; (4) enzyme systems that synthesize *prostaglandins*, which are local hormones that cause many different types of vascular and other local tissue reactions; (5) an important protein called *fibrin-stabilizing factor*, which we will discuss later in relation to blood coagulation; and (6) a *growth factor* that can cause vascular endothelial cells, vascular smooth muscle cells, and fibroblasts to multiply and grow, thus causing cellular growth that helps repair damaged vascular walls.

The cell membrane of the platelets is also important. On its surface is a coat of *glycoproteins* that causes it to adhere to injured areas of the vessel wall, especially to injured endothelial cells and even more so to any exposed collagen from deeper in the vessel wall. In addition, the membrane contains large amounts of *phospholipids* that can activate one of the blood-clotting systems, the so-called "intrinsic" system that we shall discuss later. Also, the membrane contains the protein enzyme *adenylate cyclase*, which when activated causes the formation of *cyclic AMP* inside the platelets, and this in turn excites other activities in the platelet.

Thus, the platelet is a very active structure. It has a half-life in the blood of 8 to 12 days, at the end of which time its life processes appear to run out. Then it is eliminated from the circulation mainly by the tissue macrophage system; more than half the platelets are removed by the macrophages in the spleen where the blood passes through a latticework of tight trabeculae

Mechanism of the Platelet Plug. Platelet repair of vascular openings is based on several important functions of the platelet itself. When platelets come in contact with a damaged vascular surface, such as the collagen fibers in the vascular wall or even damaged endothelial cells, they immediately change their characteristics drastically. They begin to swell; they assume irregular forms with numerous irradiating processes protruding from their surfaces; they become sticky so that they stick to the collagen fibers; and they secrete large quantities of *ADP*, and their enzymes form *thromboxane A*, a type of prostaglandin, that is also secreted by the platelets into the blood. The ADP and thromboxane A, in turn, act on nearby platelets to activate them as well, and the stickiness of these additional platelets causes them to adhere to the originally activated platelets. Therefore, at the site of any rent in a vessel, the damaged vascular wall or extravascular tissues elicit a vicious cycle of activation of successively increasing numbers of platelets; these accumulate to form a *platelet plug*. This is a fairly loose plug, but it is usually successful in blocking the blood loss if the vascular opening is small. Then, during the subsequent process of blood coagulation, to be described in subsequent paragraphs, *fibrin threads* form that attach to the platelets, thus forming a tight and unyielding plug.

Importance of the Platelet Method for Closing Vascular Holes. If the rent in a vessel is small, the platelet plug by itself can stop blood loss completely, but if there is a large hole, a blood clot in addition to the platelet plug is required to stop the bleeding.

The platelet-plugging mechanism is extremely important for closing the minute ruptures in very small blood vessels that occur hundreds of times daily, including those through the endothelial cells themselves. A person who has very few platelets develops literally hundreds of small hemorrhagic areas under the skin and throughout the internal tissues, but this does not occur in the normal person. The platelet-plugging mechanism usually does not occlude the vessel itself but merely plugs the hole, so that the vessel continues to function normally. This often is not true when a blood clot occurs.

BLOOD COAGULATION IN THE RUPTURED VESSEL

The third mechanism for hemostasis is formation of the blood clot. The clot begins to develop in 15 to 20 seconds if the trauma of the vascular wall has been severe and in 1 to 2 minutes if the trauma has been minor. Activator substances both from the traumatized vascular wall and from platelets and blood proteins adhering to the traumatized vascular wall initiate the clotting process. The physical events of this process are illustrated in Figure 8–1, and the chemical events will be discussed in detail later in the chapter.

Within three to six minutes after rupture of a vessel, if the vessel opening is not too large, the entire opening or broken end of the vessel is filled with clot. After 30 minutes to an hour, the clot retracts; this closes the vessel still further. Platelets play an important role in this clot retraction, as will also be discussed later in the chapter.

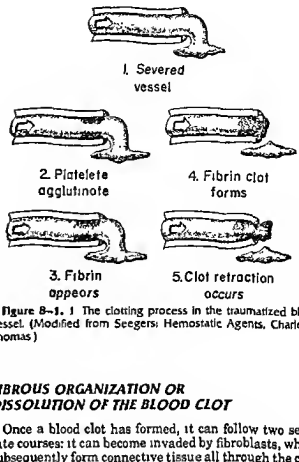


Figure 8–1. The clotting process in the traumatized blood vessel. (Modified from Seegers: Hemostatic Agents. Charles C Thomas.)

FIBROUS ORGANIZATION OR DISSOLUTION OF THE BLOOD CLOT

Once a blood clot has formed, it can follow two separate courses: it can become invaded by fibroblasts, which subsequently form connective tissue all through the clot; or it can dissolute. The usual course for a clot that forms in a small hole of a vessel wall is invasion by fibroblasts, beginning within a few hours after the clot is formed (which is promoted at least partially by the *growth factor* secreted by platelets) and continuing to complete organization of the clot into fibrous tissue within approximately seven to ten days. On the other hand, when *a large amount of blood coagulates to form one large* blood clot, such as blood that has leaked into tissues, special substances within the clot itself become activated, and these then function as enzymes to dissolute the clot itself, as will be discussed later in the chapter.

MECHANISM OF BLOOD COAGULATION

Basic Theory. Over 40 different substances that affect blood coagulation have been found in the blood and tissues, some promoting coagulation, called *procoagulants*, and others inhibiting coagulation, called *anticoagulants*. Whether or not the blood will coagulate depends on the degree of balance between these two groups of substances. Normally the anticoagulants predominate and the blood does not coagulate, but when a vessel is ruptured the activity of the procoagulants in the area of damage becomes much greater than that of the anticoagulants, and then a clot does develop.

General Mechanism. All research workers in the field of blood coagulation agree that clotting takes place in three essential steps:

First, a substance or complex of substances called *prothrombin activator* is formed in response to rupture of the vessel or damage to the blood itself.

Second, the prothrombin activator catalyzes the conversion of *prothrombin* into *thrombin*.

Third, the thrombin acts as an enzyme to convert *fibrinogen* into *fibrin threads* that enmesh platelets, blood cells, and plasma to form the clot itself.

Let us first discuss the mechanism by which the blood clot itself is formed, beginning with the conversion of prothrombin to thrombin, then we will come back to the initiating stages in the clotting process by which prothrombin activator is formed

CONVERSION OF PROTHROMBIN TO THROMBIN

After prothrombin activator has been formed as a result of rupture of the blood vessel or as a result of damage to special activator substances in the blood itself, the prothrombin activator can then cause conversion of prothrombin to thrombin, which in turn causes polymerization of fibrinogen molecules into fibrin threads within another 10 to 15 seconds. Thus the rate-limiting factor in causing blood coagulation is usually the formation of prothrombin activator and not the subsequent reactions beyond that point, for these normally occur very rapidly to form the clot itself.

Prothrombin and Thrombin. Prothrombin is a plasma protein, an alpha$_2$-globulin, having a molecular weight of 68,700 It is present in normal plasma in a concentration of about 15 mg/dl It is an unstable protein that can split easily into smaller compounds, one of which is *thrombin*, which has a molecular weight of 33,700, almost exactly half that of prothrombin

Promthrombin is formed continually by the liver, and it is continually being used throughout the body for blood clotting If the liver fails to produce prothrombin, its concentration in the plasma falls too low to provide normal blood coagulation within 24 hours. Vitamin K is required by the liver for normal formation of prothrombin as well as three other clotting factors we shall discuss later; therefore, either lack of vitamin K or the presence of liver disease that prevents normal prothrombin formation can often decrease the prothrombin level so low that a bleeding tendency results.

Effect of Prothrombin Activator on Forming Thrombin from Prothrombin. Figure 8-2 illustrates the conversion of prothrombin to thrombin under the influence of prothrombin activatar and calcium ions The rate of formation of thrombin from prothrombin is almost directly proportional to the quantity of prothrombin activator available, which itself is approximately proportional to the degree of trauma to the vessel wall or to the blood In turn, the rapidity of the clotting process is proportional to the quantity of thrombin formed

CONVERSION OF FIBRINOGEN TO FIBRIN—FORMATION OF THE CLOT

Fibrinogen. Fibrinogen is a high molecular weight protein (340,000) occurring in the plasma in quantities of 100 to 700 mg/dl Fibrinogen is formed in the liver, and liver disease occasionally decreases the concentration of circulating fibrinogen, as it does the concentration of prothrombin as was pointed out previously.

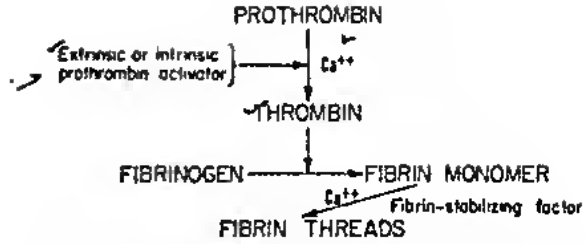


Figure 8-2. Schema for conversion of prothrombin to thrombin and polymerization of fibrinogen to form fibrin threads

Because of its large molecular size, very little fibrinogen normally leaks into the interstitial fluids, and since it is one of the essential factors in the coagulation process, interstitial fluids ordinarily coagulate poorly if at all Yet, when the permeability of the capillaries *becomes pathologically increased, fibrinogen does then* appear in the tissue fluids in sufficient quantities to allow clotting in much the same way that plasma and whole blood clot

Action of Thrombin on Fibrinogen to Form Fibrin. Thrombin is a protein *enzyme* with proteolytic capabilities It acts on fibrinogen to remove four low molecular weight peptides from each molecule of fibrinogen, forming a molecule of *fibrin monomer* that has the automatic capability of polymerizing with other fibrin monomer molecules. Therefore, many fibrin monomer molecules polymerize within seconds into *long fibrin threads* that form the *reticulum* of the clot.

In the early stages of this polymerization, the fibrin monomer molecules are held together by weak noncovalent bonds, and the threads also are not cross-linked with each other; therefore, the resultant clot is weak and can be broken apart with ease. However, still another process occurs during the following few minutes that greatly strengthens the fibrin reticulum This involves a substance called *fibrin-stabilizing factor* that is normally present in small amounts in the plasma globulins but that is also released from platelets entrapped in the clot Before fibrin-stabilizing factor can have an effect on the fibrin threads it must itself be activated. Fortunately, the same thrombin that causes fibrin formation also activates the fibrin-stabilizing factar. Then this activated substance operates as an enzyme to cause *covalent* bonds between the fibrin monomer molecules as well as multiple cross-linkages between the adjacent fibrin threads, thus adding tremendously to the three-dimensional strength of the fibrin meshwork.

The Blood Clot. The clot is composed of a meshwork of fibrin threads running in all directions and entrapping blood cells, platelets, and plasma The fibrin threads adhere to damaged surfaces of blood vessels, therefore, the blood clot becomes adherent to any vascular opening and thereby prevents blood loss

Clot Retraction—Serum. Within a few minutes after a clot is formed, it begins to contract and usually expresses most of the fluid from the clot within 30 to 60 minutes The fluid expressed is called *serum*, because all its fibrinogen and most of the other clotting factors have been removed, in this way, serum differs from plasma. Serum obviously cannot clot because of lack of these factors.

Platelets are necessary for clot retraction to occur. Therefore, failure of clot retraction is an indication that

the number of platelets in the circulating blood is low. Electron micrographs of platelets in blood clots show that they become attached to the fibrin threads in such a way that they actually bond different threads together. Furthermore, platelets entrapped in the clot continue to release procoagulant substances, one of which is fibrin-stabilizing factor that causes more and more cross-linking bonds between the adjacent fibrin threads. In addition, the platelets themselves contribute directly to clot contraction by activating the platelet actin and myosin molecules, which are contractile proteins and cause strong contraction of the platelet spicules attached to the fibrin. Obviously, this also helps compress the fibrin meshwork into a smaller mass. The contraction of the actin and myosin molecules is probably activated by thrombin as well as by calcium ions released from the calcium stores in the endoplasmic reticulum and Golgi apparatus of the platelets.

As the clot retracts, the edges of the broken blood vessel are pulled together, thus possibly or probably contributing to the ultimate state of hemostasis.

THE VICIOUS CIRCLE OF CLOT FORMATION

Once a blood clot has started to develop, it normally extends within minutes into the surrounding blood. That is, the clot itself initiates a vicious circle to promote more clotting. One of the most important causes of this is the fact that the proteolytic action of thrombin allows it to act on many of the other blood-clotting factors in addition to fibrinogen. For instance, thrombin has a direct proteolytic effect on prothrombin itself, tending to split this into still more thrombin, and it acts on some of the blood-clotting factors responsible for the formation of prothrombin activator. (These effects, to be discussed in subsequent paragraphs, include (a) acceleration of the actions of factors VIII, IX, X, XI, and XII, and (b) aggregation of platelets.) Once a critical amount of thrombin is formed, a vicious circle develops that causes still more blood clotting and more thrombin to be formed; thus, the blood clot continues to grow until something stops its growth.

BLOCK OF CLOT GROWTH BY BLOOD FLOW

Fortunately, when a clot develops, the vicious circle of continued clot formation occurs only where the blood is not moving because flowing blood carries away the thrombin and the other procoagulants released during the clotting process so rapidly that their concentrations cannot rise high enough to promote further clotting. Thus, extension of the clot almost always stops where it comes in contact with blood that is flowing faster than a certain velocity.

The clotting process fortunately is not self-propagating when the concentrations of the procoagulants are below critical concentrations. At these lower concentrations, many inhibitor substances in the blood, some of which will be discussed later in the chapter, continually block the actions of the procoagulants or destroy them. In addition, the macrophage system, particularly the Kupffer cells of the liver and the macrophages of the spleen and bone marrow, removes most of the circulating procoagulants within a few minutes.

INITIATION OF COAGULATION: FORMATION OF PROTHROMBIN ACTIVATOR

Now that we have discussed the clotting process initiated by the formation of thrombin from prothrombin, we must turn to the more complex mechanisms that activate the prothrombin. These mechanisms can be set into play by trauma to the tissues, trauma to the blood, or contact of the blood with damaged endothelial cells or with collagen and other tissue elements outside the blood vessel endothelium. In each instance, they lead to the *formation of prothrombin activator*, which then causes prothrombin conversion to thrombin.

Prothrombin activator can be formed in two basic ways: (1) by the *extrinsic pathway* that begins with trauma to the vascular wall and surrounding tissues, or (2) by the *intrinsic pathway* that begins in the blood itself.

In both the extrinsic and intrinsic pathways a series of different plasma proteins, especially beta-globulins, play major roles. These, along with the other factors already discussed that enter into the clotting process, are called *blood-clotting factors* and for the most part they are inactive forms of proteolytic enzymes. When converted to the active forms, their enzymatic actions cause the successive, cascading reactions of the clotting process.

Most of the clotting factors are designated by Roman numerals, as listed in Table 8–1.

The Extrinsic Mechanism for Initiating Clotting

The extrinsic mechanism for initiating the formation of prothrombin activator begins with blood coming in contact with traumatized vascular wall or extravascular tissues and occurs according to the following three basic steps, as illustrated in Figure 8–3.

(1) *Release of tissue thromboplastin.* Traumatized tissue releases a complex of several factors called *tissue thromboplastin.* This includes especially *phospholipids*

Table 8–1 CLOTTING FACTORS IN THE BLOOD AND THEIR SYNONYMS

Clotting Factor	Synonyms
Fibrinogen	Factor I
Prothrombin	Factor II
Tissue thromboplastin	Factor III, tissue factor
Calcium	Factor IV
Factor V	Proaccelerin; labile factor, Ac-globulin, Ac-G
Factor VII	Serum prothrombin conversion accelerator, proconvertin; SPCA; stable factor
Factor VIII	Antihemophilic factor; AHF; antihemophilic globulin; AHG, antihemophilic factor A
Factor IX	Plasma thromboplastin component, PTC, Christmas factor, antihemophilic factor B
Factor X	Stuart factor, Stuart-Prower factor
Factor XI	Plasma thromboplastin antecedent, PTA; antihemophilic factor C
Factor XII	Hageman factor
Factor XIII	Fibrin-stabilizing factor
Platelets	

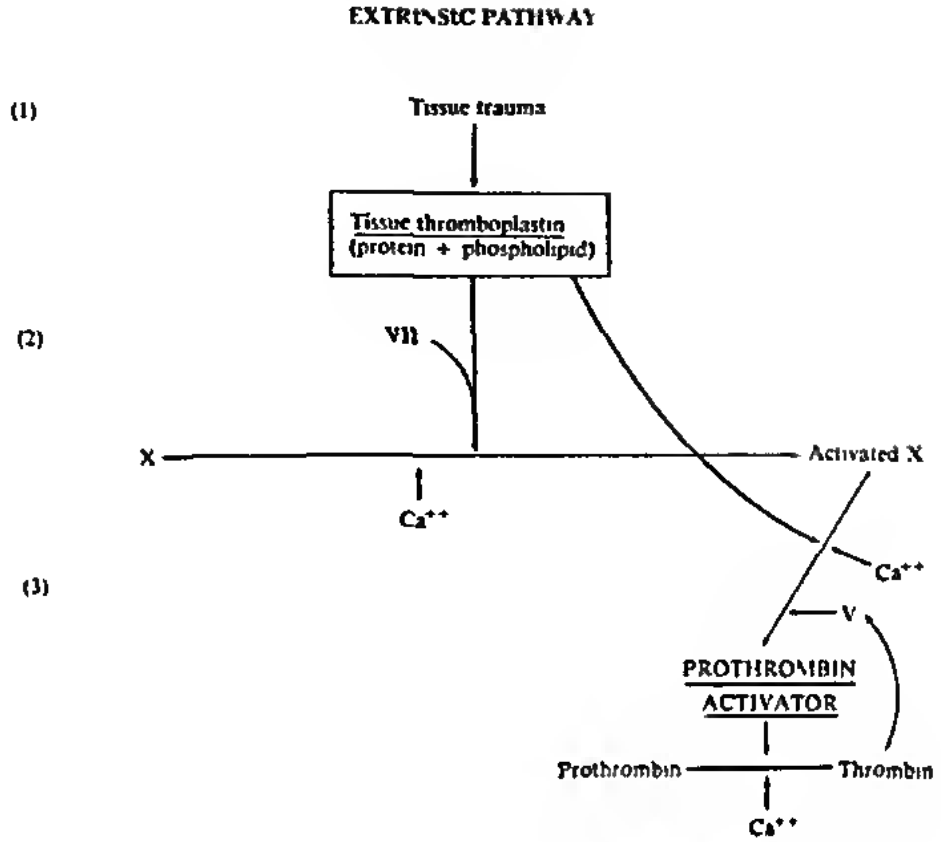


Figure 8–3. The extrinsic pathway for initiating blood clotting.

from the membranes of the tissues and at least one important *glycoprotein*, which functions as a *proteolytic enzyme*

(2) *Activation of factor X to form activated factor X—role of factor VII and tissue factor* The tissue glycoprotein complexes with blood coagulation factor VII, and this complex, in the presence of tissue phospholipids, acts enzymatically on factor X to form *activated factor X*

(3) *Effect of activated factor X to form prothrombin activator—role of factor V* The activated factor X complexes immediately with the tissue phospholipids released as part of the tissue thromboplastin and also with factor V to form the complex called *prothrombin activator* Within a few seconds this splits prothrombin to form thrombin, and the clotting process proceeds as has already been explained At first, the factor V in the prothrombin activator complex is inactive, but once clotting begins the proteolytic action of thrombin activates factor V, this then becomes an additional strong accelerator of prothrombin activation Thus, in the final prothrombin activator complex, activated factor X is the actual protease that causes splitting of prothrombin to thrombin, activated factor V greatly accelerates this protease activity, and phospholipids act as a vehicle that further accelerates the process Note especially the *positive feedback* effect of thrombin, acting through factor V, in accelerating the entire process once it begins.

The Intrinsic Mechanism for Initiating Clotting

The second mechanism for initiating the formation of prothrombin activator, and therefore for initiating clotting, begins with trauma to the blood itself and continues through the following series of cascading reactions, as illustrated in Figure 8–4

(1) *Activation of factor XII and release of platelet phospholipids by blood trauma.* Trauma to the blood alters two important clotting factors in the blood—factor XII and the platelets. When factor XII is disturbed, such as by coming into contact with collagen or with a wettable surface such as glass (particularly a negatively charged surface), it takes on a new configuration that converts it into a proteolytic enzyme called "activated factor XII"

Simultaneously, the blood trauma also damages the platelets, either because of adherence to collagen or to a wettable surface (or by damage in other ways), and this releases platelet phospholipid which also plays a role in subsequent clotting reactions.

(2) *Activation of factor XI* The activated factor XII acts enzymatically on factor XI to activate this as well, which is the second step in the intrinsic pathway. This reaction also requires *HMW kinogen*, and it is accelerated by prekallikrein.

(3) *Activation of factor IX by activated factor XI* The activated factor XI then acts enzymatically on factor IX to activate this factor, also.

(4) *Activation of factor X—role of factor VIII.* The activated factor IX, acting in concert with factor VIII and with the platelet phospholipids from the traumatized platelets, activates factor X. It is clear that when either factor VIII or platelets are in short supply, this step is deficient. Factor VIII is the factor that is missing in the person who has classic *hemophilia*, for which reason it is called *antihemophilic factor*. Platelets are the clotting factor that is lacking in the bleeding disease called *thrombocytopenia*.

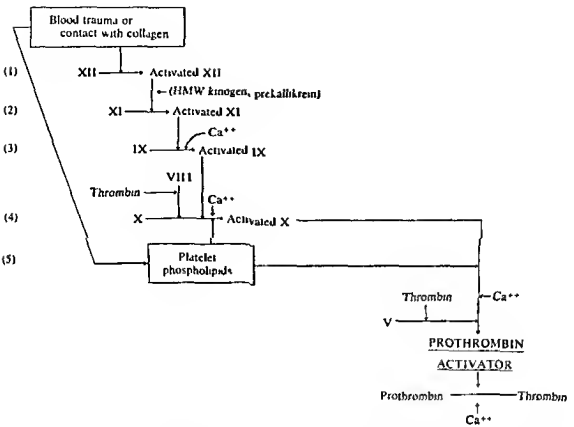


Figure 8–4. The intrinsic pathway for initiating blood clotting.

(5) *Action of activated factor X to form prothrombin activator—role of factor V.* This step in the intrinsic pathway is essentially the same as the last step in the extrinsic pathway. That is, activated factor X combines with factor V and platelet phospholipids to form the complex called *prothrombin activator.* The only difference is that the phospholipids in this instance come from the traumatized platelets rather than from traumatized tissues. The prothrombin activator in turn initiates within seconds the cleavage of prothrombin to form thrombin, thereby setting into motion the final clotting process, as described earlier.

Role of Calcium Ions in the Intrinsic and Extrinsic Pathways

Except for the first two steps in the intrinsic pathway, calcium ions are required for promotion of all the reactions. Therefore, in the absence of calcium ions, blood clotting will not occur.

Fortunately, in the living body the calcium ion concentration rarely falls low enough to affect significantly the kinetics of blood clotting. The reason for this is that before calcium ion concentration can fall this low, the diminished level of calcium ions will likely kill the person by causing muscle tetany throughout the body, especially of the respiratory muscles.

On the other hand, when blood is removed from a person, it can be prevented from clotting by reducing *the calcium ion concentration below the threshold level* for clotting, either by deionizing the calcium by reacting it with substances such as *citrate ion* or by precipitating the calcium with substances such as *oxalate ion.*

Summary of Blood Clotting Initiation

It is clear from the above schemas of the intrinsic and extrinsic systems for initiating blood clotting that clotting is initiated after rupture of blood vessels by both the pathways. The tissue thromboplastin initiates the extrinsic pathway, whereas contact of factor XII and the platelets with collagen in the vascular wall initiates the intrinsic pathway

In contrast, when blood is removed from the body and held in a test tube, the intrinsic pathway alone must elicit the clotting. This usually results from contact of factor XII and platelets with the wall of the vessel, which activates both of these and initiates the intrinsic mechanism. If the surface of the container is very "nonwettable," such as a siliconized surface, blood clotting can sometimes be prevented for an hour or more.

Intravascular clotting sometimes results also from other factors that activate the intrinsic pathway. For instance, antigen-antibody reactions sometimes initiate the clotting process, and the same is true of some drugs or debris that might happen to enter the circulation.

An especially important difference between the extrinsic and intrinsic pathways is that *the extrinsic pathway* can be explosive in nature; once initiated, its speed of occurrence is limited only by the amount of tissue thromboplastin released from the traumatized tissues, and by the quantities of factors X, VII, and V in the blood. With severe tissue trauma, clotting can occur in as little as 15 seconds. On the other hand, the intrinsic *pathway is much slower to proceed,*
two to six minutes to cause
inhibitors in the blood set up readily
for the intrinsic pathway. At times,

block the intrinsic pathway entirely; some of them will be discussed in more detail later in the chapter.

PREVENTION OF BLOOD CLOTTING IN THE NORMAL VASCULAR SYSTEM—THE INTRAVASCULAR ANTICOAGULANTS

Endothelial Surface Factors. Probably the two most important factors for preventing clotting in the normal vascular system are, first, the smoothness of the endothelium that prevents contact activation of the intrinsic clotting system and, second, a monomolecular layer of protein adsorbed to the inner surface of the endothelium that repels the clotting factors and platelets, thereby preventing activation of clotting. When the endothelial wall is damaged, its smoothness and protein layer are both lost, which activates both factor XII and the platelets, thus setting off the intrinsic pathway of clotting. And, if factor XII and platelets come in contact with the subendothelial collagen, the activation is even more powerful.

The Antithrombin Action of Fibrin and of Antithrombin III. Among the most important anticoagulants in the blood itself are those that remove thrombin from the blood. The two most powerful of these are the *fibrin threads* formed during the process of clotting and an alpha-globulin called *antithrombin III* or also *antithrombin-heparin cofactor*.

While a clot is forming, approximately 85 to 90 per cent of the thrombin formed from the prothrombin becomes adsorbed to the fibrin threads as they develop. This obviously helps prevent the spread of thrombin into the remaining blood and therefore prevents excessive spread of the clot.

The thrombin that does not adsorb to the fibrin threads soon combines with antithrombin III, which blocks the effect of the thrombin on the fibrinogen and then inactivates the bound thrombin during the next 12 to 20 minutes.

Heparin. Heparin is another powerful anticoagulant. Yet, its concentration in the blood is normally very slight, so that only under some physiological conditions does it have significant anticoagulant effects. On the other hand, it is used very widely in medical practice to prevent intravascular clotting.

The heparin molecule is a highly negatively charged conjugated polysaccharide. By itself it has little or no anticoagulant property, but it combines with antithrombin III and increases as much as a thousandfold the effectiveness of antithrombin III to remove thrombin and thus to act as an anticoagulant. Therefore, in the presence of excess heparin the removal of thrombin from the circulating blood is almost instantaneous.

The complex of heparin and antithrombin III also removes several other activated coagulation factors in addition to thrombin, thus further enhancing the effectiveness of anticoagulation. These others include factors XII, XI, IX, and X.

Heparin is produced by many different cells of the human body, though especially large quantities are formed by the basophilic *mast cells* located in the pericapillary connective tissue throughout the body. These cells continually secrete small quantities of heparin, and minute quantities also diffuse into the circulatory system. The *basophil cells* of the blood, which seem to be functionally almost identical with the mast cells, also release small quantities of heparin into the plasma.

Mast cells are extremely abundant in the tissue surrounding the capillaries of the lungs and to a lesser extent the capillaries of the liver. It is easy to understand why large quantities of heparin might be needed in these areas, for the capillaries of the lungs and liver receive many embolic clots formed in the slowly flowing venous blood; sufficient formation of heparin might prevent further growth of the clots.

Alpha$_2$-Macroglobulin. *Alpha$_2$-macroglobulin* is a very large globulin molecule having a molecular weight of 360,000. It is similar to antithrombin-heparin cofactor in that it combines with the proteolytic coagulation factors. However, its activity is not accelerated by heparin. Its function is mainly to act as a binding agent for the coagulation factors and prevent their proteolytic actions until they can be destroyed in various ways, but not by the alpha$_2$-macroglobulin itself. In this way it probably plays an important role even normally in preventing blood clotting.

LYSIS OF BLOOD CLOTS—PLASMIN

The plasma proteins contain a euglobulin called *plasminogen* or *profibrinolysin* which, when activated, becomes a substance called *plasmin* or *fibrinolysin*. Plasmin is a proteolytic enzyme that resembles trypsin, the most important digestive enzyme of pancreatic secretion. It digests the fibrin threads and also digests other substances in the surrounding blood, such as fibrinogen, factor V, factor VIII, prothrombin, and factor XII. Therefore, whenever plasmin is formed in a blood clot, it can cause lysis of the clot and also destruction of many of the clotting factors, thereby causing hypocoagulability of the blood.

Formation of Plasmin and Lysis of Clots. When a clot is formed a large amount of plasminogen is incorporated in the clot along with other plasma proteins. However, this will not become plasmin and will not cause lysis of the clot until it is activated. Fortunately, the tissues and blood contain substances that can activate plasminogen to plasmin, including (1) thrombin, (2) activated factor XII, (3) lysosomal enzymes from damaged tissues, and (4) factors from the vascular endothelium. Within a day or two after blood has leaked into a tissue and clotted, these activators usually cause the formation of enough plasmin that it in turn begins to dissolve the clot.

Clots that occur inside blood vessels can also be dissolved, though this occurs more readily in the small vessels than in the large ones. This illustrates that activator systems also occur within the blood itself, especially the thrombin formed during the clotting process.

An activator called *urokinase* is found in the urine and is believed to be important in the lysis of clots that develop in the renal tubules. It is possible that urokinase also plays a role as an intravascular activator before it is excreted into the urine by the kidneys.

Certain bacteria also release activator enzymes; for instance, streptococci release a substance called *streptokinase*, which acts on plasminogen to form plasmin. When streptococcal infection occurs in the tissues, the plasmin generated in response to the streptokinase dissolves clotted lymph and clotted tissue fluids and allows the streptococci to spread extensively through the tissues instead of being blocked by the body's "walling off" process described in Chapter 5.

A Plasmin Inhibitor, α_2-Antiplasmin. Plasmin not only destroys fibrin threads but also functions as a proteolytic enzyme to digest fibrinogen and a number of other clotting factors as well. Small amounts of plasmin are formed in the blood all the time, which could seriously impede the activation of the clotting system were it not for the fact that the blood also contains another factor, α_2-*antiplasmin*, that binds with plasmin and inhibits it. Therefore, the rate of plasmin formation must rise above a certain critical level before it becomes effective.

Significance of the Fibrinolysis System. The lysis of blood clots allows slow clearing (over a period of several days) of extraneous blood in the tissues and sometimes allows reopening of clotted vessels. Unfortunately, reopening of large vessels occurs only rarely. But an important function of the fibrinolysis system is to remove very minute clots from the millions of tiny peripheral vessels that eventually would all become occluded were there no way to cleanse them.

CONDITIONS THAT CAUSE EXCESSIVE BLEEDING IN HUMAN BEINGS

Excessive bleeding can result from deficiency of any one of the many different blood clotting factors. Three particular types of bleeding tendencies that have been studied to the greatest extent will be discussed: (1) bleeding caused by vitamin K deficiency, (2) hemophilia, and (3) thrombocytopenia (platelet deficiency).

DECREASED PROTHROMBIN, FACTOR VII, FACTOR IX, AND FACTOR X CAUSED BY VITAMIN K DEFICIENCY

With few exceptions, almost all the blood-clotting factors are formed by the liver. Therefore, diseases of the liver such as *hepatitis*, *cirrhosis*, and *acute yellow atrophy* all can depress the clotting system so greatly that the patient develops a severe tendency to bleed.

Another cause of depressed formation of clotting factors by the liver is vitamin K deficiency. Vitamin K is necessary to promote the formation of four of the most important clotting factors, *prothrombin*, *factor VII*, *factor IX*, and *factor X*. In the absence of vitamin K, insufficiency of these coagulation factors can also lead to a serious bleeding tendency.

Fortunately, vitamin K is continually synthesized in the intestinal tract by bacteria so that vitamin K deficiency rarely if ever occurs simply because of its absence from the diet, except in newborn children before they establish their intestinal bacterial flora. However, vitamin K deficiency does often occur as a result of poor absorption of fats from the gastrointestinal tract, because vitamin K is fat-soluble and ordinarily is absorbed into the blood along with the fats.

One of the most prevalent causes of vitamin K deficiency is failure of the liver to secrete bile into the gastrointestinal tract (which occurs either as a result of obstruction of the bile ducts or as a result of liver disease), for lack of bile prevents adequate fat digestion and absorption. Therefore, liver disease often causes decreased production of prothrombin and the other factors both because of poor vitamin K absorption and because of dysfunctional liver cells. Because of this, vitamin K is injected into all patients with liver disease or obstructed bile ducts prior to performing any surgical procedure. Ordinarily, if vitamin K is given to a deficient patient four to eight hours prior to operation and the liver parenchymal cells are at least one-half normal in function, sufficient clotting factors will be produced to prevent excessive bleeding during the operation.

HEMOPHILIA

Hemophilia is a bleeding tendency that occurs almost exclusively in males. In 85 per cent of cases it is caused by *deficiency of factor VIII*; this type of hemophilia is called *hemophilia A* or *classic hemophilia*. About one of every 10,000 males in the United States has classic hemophilia. In the other 15 per cent of the patients, the bleeding tendency is caused by deficiency of factor IX. Both these factors are transmitted genetically by way of the female chromosome as a recessive trait. Therefore, almost never will a woman have hemophilia because at least one of her two X chromosomes will have the appropriate genes. However, if one of her X chromosomes is deficient, she will be a *hemophilia carrier*, transmitting the disease to half her male offspring and transmitting the carrier state to half her female offspring.

The bleeding trait in hemophilia can have various degrees of severity, depending upon the severity of the genetic deficiency. Usually, bleeding will not occur except after trauma, but the degree of trauma required to cause severe and prolonged bleeding may be so mild that it is hardly noticeable. For instance, even falling on a knee may cause severe bleeding into the knee joint. Likewise, a mild blow to the head can frequently cause intracranial bleeding, or bleeding can occur in fascial compartments in the arm or leg and compress both nerves and blood vessels to the extent of causing severe atrophy of the limbs. And bleeding can often last for literally weeks after extraction of a tooth.

The blood clotting factor VIII is composed of two separate components, a very large component with a molecular weight in the millions and another smaller component with a molecular weight of about 230,000. This smaller component is most important in the intrinsic pathway for clotting, and it is deficiency of this part of factor VIII that causes classic hemophilia. Another bleeding disease with somewhat different characteristics, called von Willebrand's disease, results from loss of the large component.

When a person with classic hemophilia develops severe, prolonged bleeding, almost the only therapy that is truly effective is injection of purified clotting factor VIII. This usually relieves the person's bleeding tendency for several days, and blood vessel repair is often attained during this period of time. Unfortunately, the cost of factor VIII is very high, and its availability is limited because it can be gathered only from human blood and only in extremely small quantities.

THROMBOCYTOPENIA

Thrombocytopenia means the presence of a very low quantity of platelets in the circulatory system. Persons with thrombocytopenia have a tendency to bleed as do hemophiliacs, except that the bleeding is usually from many small capillaries rather than from larger vessels,

as in hemophilia. As a result, small punctate hemorrhages occur throughout all the body tissues. The skin of such a person displays many small, purplish blotches, giving the disease the name *thrombocytopenic purpura*. It will be remembered that platelets are especially important for repair of minute breaks in capillaries and other small vessels. Indeed, platelets can aggregate to fill such ruptures without actually causing clots.

Ordinarily, bleeding does not occur until the number of platelets in the blood falls below a value of approximately 50,000 per cubic millimeter rather than the normal 150,000 to 350,000. Levels as low as 10,000 per cubic millimeter are frequently lethal.

Even without making specific platelet counts on the blood, one can sometimes suggest the existence of thrombocytopenia by simply noting whether or not a clot of the person's blood retracts, for, as pointed out earlier, clot retraction is normally dependent upon the presence of large numbers of platelets entrapped in the fibrin mesh of the clot.

Most persons with thrombocytopenia have the disease known as *idiopathic thrombocytopenia*, which means simply "thrombocytopenia of unknown cause." However, in the past few years it has been discovered that in most of these persons specific antibodies are destroying the platelets. Occasionally these have developed because of transfusions from other persons, but usually they result from development of autoimmunity to the person's own platelets, the cause of which, however, is not known.

In addition to idiopathic thrombocytopenia, the number of thrombocytes (platelets) in the blood may be greatly depressed by any abnormality that causes aplasia of the bone marrow. For instance, *irradiation injury* to the bone marrow, aplasia of the bone marrow resulting from *drug sensitivity*, and even *pernicious anemia* can cause sufficient decrease in the total number of platelets that thrombocytopenic bleeding results.

Relief from bleeding for one to four days can often be effected in the thrombocytopenic patient by giving *fresh whole blood transfusions*. To do this, the blood is best removed from the donor into a siliconized chamber and then rapidly placed in the recipient so that the platelets are damaged as little as possible. Prednisone and azathioprine, which suppress the formation of antibodies, are often beneficial in the idiopathic type of thrombocytopenia. And, *splenectomy* is usually very helpful, sometimes effecting almost a complete cure, because the spleen removes large numbers of platelets, particularly damaged ones, from the blood.

THROMBOEMBOLIC CONDITIONS IN THE HUMAN BEING

Thrombi and Emboli. An abnormal clot that develops in a blood vessel is called a *thrombus*. Once a clot has developed, continued flow of blood past the clot is likely to break it away from its attachment, and such freely flowing clots are known as *emboli*. Emboli generally do not stop flowing until they come to a narrow point in the circulatory system. Thus, emboli originating in large arteries or in the left side of the heart eventually plug either smaller systemic arteries or arterioles, maybe in the brain, the kidneys, or elsewhere. On the other hand, emboli originating in the venous system and in the right side of the heart flow into the vessels of the lung to cause pulmonary arterial embolism.

Causes of Thromboembolic Conditions. The causes of thromboembolic conditions in the human being are usually twofold. First, any *roughened endothelial surface of a vessel*—as may be caused by arteriosclerosis, infection, or trauma—is likely to initiate the clotting process. Second, blood often clots *when it flows very slowly* through blood vessels, for small quantities of thrombin and other procoagulants are always being formed. These are generally removed from the blood by the macrophage system, mainly the Kupffer cells of the liver. If the blood is flowing too slowly, the concentrations of the procoagulants in local areas often rise high enough to initiate clotting, but when the blood flows rapidly these are rapidly mixed with large quantities of blood and are removed during passage through the liver.

FEMORAL THROMBOSIS AND MASSIVE PULMONARY EMBOLISM

Because clotting almost always occurs when blood flow is blocked for many hours in any vessel of the body, the immobility of bed patients plus the practice of propping the knees with underlying pillows often causes intravascular clotting because of blood stasis in one or more of the leg veins for hours at a time. Then the clot grows, mainly in the direction of the slowly moving blood, sometimes growing the entire length of the leg veins and occasionally even up into the common iliac vein and inferior vena cava. Then, about one time out of every ten, a large part of the clot disengages from its attachments to the vessel wall and flows freely with the venous blood into the right side of the heart and thence into the pulmonary arteries to cause *massive pulmonary embolism*. If the clot is large enough to occlude both the pulmonary arteries, immediate death ensues. If only one pulmonary artery or a smaller branch is blocked, death may not occur, or the embolism may lead to death a few hours to several days later because of further growth of the clot within the pulmonary vessels.

DISSEMINATED INTRAVASCULAR CLOTTING

Occasionally, the clotting mechanism becomes activated in widespread areas of the circulation, giving rise to the condition called *disseminated intravascular clotting*. This often results from the presence of large amounts of traumatized or dying tissue in the body that releases tissue thromboplastin into the blood. Frequently, the clots are small but numerous, and they plug a large share of the small peripheral blood vessels. This occurs especially in septicemic shock, in which either circulating bacteria or bacterial toxins—especially *endotoxins*—activate the clotting mechanisms. The plugging of the small peripheral vessels greatly diminishes the delivery of oxygen and other nutrients to the tissues—a situation which exacerbates the shock picture. It is partly for this reason that full-blown septicemic shock is lethal in 85 per cent or more of the patients.

A peculiar effect of disseminated intravascular clotting is that the patient frequently begins to bleed. The reason for this is that so many of the clotting factors are removed by the widespread clotting that too few procoagulants remain to allow normal hemostasis of the remaining blood.

ANTICOAGULANTS FOR CLINICAL USE

In some thromboembolic conditions it is desirable to delay the coagulation process to a certain degree. Therefore, various anticoagulants have been developed for treatment of these conditions. The ones most useful clinically are heparin and the coumarins.

HEPARIN AS AN INTRAVENOUS ANTICOAGULANT

Commercial heparin is extracted from animal tissues from several organs and is prepared in almost pure form. Injection of relatively small quantities, approximately 0.5 to 1 mg per kilogram of body weight, causes the blood clotting time to increase from a normal of approximately 6 minutes to 30 or more minutes. Furthermore, this change in clotting time occurs instantaneously, thereby immediately preventing further development of the thromboembolic condition

The action of heparin lasts approximately three to four hours. It is believed that the injected heparin is destroyed by an enzyme in the blood known as *heparinase*.

In the treatment of a patient with heparin, too much heparin is sometimes given, and *serious* bleeding crises occur. In these instances, *protamine* acts specifically as an antiheparin, and the clotting mechanism can be reverted to normal by administering this substance. Protamine combines with heparin and inactivates it because it carries strong positive electrical charges, whereas heparin carries strong negative charges.

COUMARINS AS ANTICOAGULANTS

When a coumarin, such as *warfarin*, is given to a patient, the plasma levels of prothrombin and factors VII, IX, and X, all formed by the liver, begin to fall, indicating that warfarin has a potent depressant effect on liver formation of all these compounds. Warfarin causes this effect by competing with vitamin K for reactive sites in the enzymatic processes for formation of prothrombin and the other three clotting factors, thereby blocking the action of vitamin K.

After administration of an effective dose of warfarin, the coagulant activity of the blood decreases to approximately 50 per cent of normal by the end of 12 hours and to approximately 20 per cent of normal by the end of 24 hours. In other words, the coagulation process is not blocked immediately, but must await the consumption of the prothrombin and other factors already present in the plasma. Normal coagulation returns one to three days after discontinuing therapy.

PREVENTION OF BLOOD COAGULATION OUTSIDE THE BODY

Though blood removed from the body and held in a glass test tube normally clots in about six minutes, blood collected in *siliconized containers* often does not clot for as long as an hour or more. The reason for this delay is that preparing the surfaces of the containers with silicone prevents contact activation of factor XII, which initiates the intrinsic clotting mechanism. On the other hand, untreated glass containers allow contact activation and rapid development of clots.

Heparin can be used for preventing coagulation of blood outside the body as well as in the body, and heparin is occasionally used as an anticoagulant when blood is removed from a donor to be transfused later into a recipient Also, heparin is used in all surgical procedures in which the blood is passed through a heart-lung machine or artifical kidney and then back into the person

Various substances that *decrease the concentration of calcium ions* in the blood can be used for preventing blood coagulation outside the body For instance, soluble *oxalate* compounds mixed in very small quantity with a sample of blood cause precipitation of calcium oxalate from the plasma and thereby decrease the ionic calcium levels so much that blood coagulation is blocked.

A second calcium deionizing agent used for preventing coagulation is *sodium, ammonium*, or *potassium citrate* The citrate ion combines with calcium in the blood to cause an un-ionized calcium compound, and the lack of ionic calcium prevents coagulation Citrate anticoagulants have a very important advantage over the oxalate anticoagulants, for oxalate is toxic to the body, whereas moderate quantities of citrate can be injected intravenously After injection, the citrate ion is removed from the body within a few minutes by the liver and is polymerized into glucose or metabolized directly for energy. Consequently, 500 ml of blood that has been rendered incoagulable by citrate can ordinarily be injected into a recipient within a few minutes without any dire consequences If the liver is damaged or if large quantities of citrated blood or plasma are given too rapidly, the citrate ion may not be removed quickly enough, and the citrate can then greatly depress the level of calcium ion in the blood, which results in tetany and convulsive death

BLOOD COAGULATION TESTS

BLEEDING TIME

When a sharp knife is used to pierce the tip of the finger or lobe of the ear, bleeding ordinarily lasts three to six minutes. However, the time depends largely on the depth of the wound and on the degree of hyperemia in the finger at the time of the test Lack of several of the clotting factors can prolong the bleeding time, but this is especially prolonged by lack of platelets

CLOTTING TIME

Many methods have been devised for determining clotting times. The one most widely used is to collect *blood in a chemically clean glass test tube* and then to tip the tube back and forth approximately every 30 seconds until the blood has clotted. By this method, the normal clotting time is about five to eight minutes.

Procedures using multiple test tubes have been devised for determining clotting time more accurately. However, clotting times are also very dependent on the condition of the glass itself and even on the size of the tube, which makes a high degree of standardization necessary to obtain accurate results. A typical condition that causes a prolonged clotting time is hemophilia, but

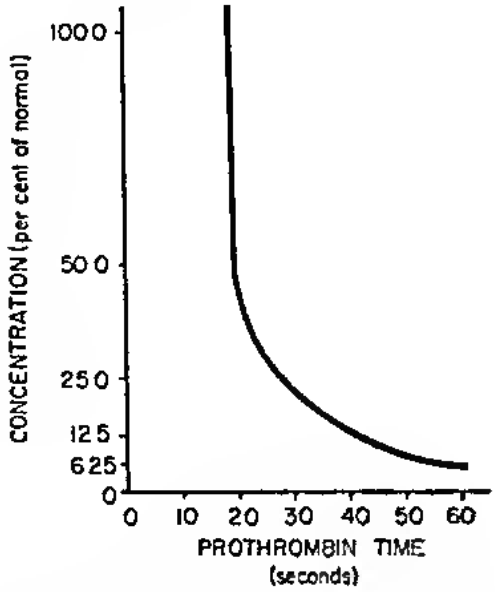


Figure 8–5. Relationship of prothrombin and concentration in the blood to the prothrombin time

deficiency of any of the factors in the intrinsic pathway for clotting can be the culprit.

PROTHROMBIN TIME

The prothrombin time gives an indication of the total quantity of prothrombin in the blood. Figure 8–5 shows the relationship of prothrombin concentration to prothrombin time. The means for determining prothrombin time is the following:

Blood removed from the patient is immediately oxalated so that none of the prothrombin can change into thrombin. At any time later, a large excess of calcium ion and tissue thromboplastin is suddenly mixed with the oxalated blood. The calcium nullifies the effect of the oxalate, and the tissue thromboplastin activates the prothrombin-to-thrombin reaction by means of the extrinsic clotting pathway. The time required for coagulation to take place is known as the "prothrombin time." The normal prothrombin time is approximately 12 seconds, though this depends to a certain extent on the exact procedure employed. In each laboratory a curve relating prothrombin concentration to prothrombin time, such as that shown in Figure 8–5, is ordinarily drawn for the method used so that the significance of the prothrombin time can be evaluated.

Tests similar to that for prothrombin time have been devised to determine the relative quantities of other clotting factors in the body. In performing the tests excesses of calcium ions and all the other factors besides the one being tested are added to oxalated blood all at once, and then the time of coagulation is determined in the same manner as the usual prothrombin time. If the factor is deficient, the time will be considerably prolonged.

REFERENCES

Biggs, R. (ed.) The Treatment of Haemophilia A and B and Von Willebrand's Disease. Philadelphia, J. B. Lippincott Co., 1978.

Carpentier, A. Low molecular weight (LMW) heparin derivatives in experimental extra-corporeal circulation (ECC). *Haemostasis, 14*:325, 1984.

Chavin, S. I. Factor VIII: Structure and function in blood clotting. *Am. J. Hematol., 16*:297, 1984.

Collen, D. *et al.* (eds.): The Physiological Inhibitors of Blood Coagulation and Fibrinolysis. New York, Elsevier/North Holland, 1979.

Engelbert, H. Heparin. New York, S. Karger, 1978.

Frojmovic, M. M. and Milton, J. G. Human platelet size, shape, and related functions in health and disease. *Physiol. Rev., 62*:185, 1982.

Fukada, E. *et al.* Methods to study rheological properties of blood during clotting. *Biorheology*, Suppl. *1*:9, 1984.

Gowans, J. L. (in honour of): Blood Cells and Vessel Walls. Functional Interactions. Princeton, N.J., Excerpta Medica, 1980.

Hassan, H. J. *et al.* Intrafemic Factor IX restriction site polymorphism in hemophilia B variants. *Blood, 65*:441, 1985.

Hirsh, J. *et al.* Concepts in Hemostasis and Thrombosis. New York, Churchill Livingstone, 1979.

Joist, J. H., and Sherman, L. A. (eds.): Venous and Atrial Thrombosis: Pathogenesis, Diagnosis, Prevention and Therapy. New York, Grune & Stratton, 1979.

Kane, K. K. Fibrinolysis—a review. *Ann. Clin. Lab. Sci., 14*:443, 1984.

Kline, D. L. and Reddy, K. N. N. (eds.): Fibrinolysis. Boca Raton, Fla., CRC Press, 1980.

Kumada, T., and Abika, Y. Purification and some properties of rat alpha 2 plasmin inhibitor. *Thromb. Res., 36*:143, 1984.

Lewis, J. H. *et al.* Bleeding Disorders. Garden City, N.Y., Medical Examination Publishing Co., 1979.

Mann, K. G. The biochemistry of coagulation. *Clin. Lab. Med., 4*:207, 1984.

Markwardt, F. (ed.): Fibrinolytics and Antifibrinolytics. New York, Springer-Verlag, 1978.

McDuffie, N. M. (ed.): Heparin: Structure, Cellular Functions, and Clinical Application. New York, Academic Press, 1979.

Murano, G., and Bick, R. L. (eds.): Basic Concepts of Hemostasis and Thrombosis: Clinical Laboratory Evaluation of Thrombohemorrhagic Phenomena. Boca Raton, Fla., CRC Press, 1980.

Mustard, J. F. *et al.* Prostaglandins and platelets. *Annu. Rev. Med., 31*:89, 1980.

Nemerson, Y., and Nossel, H. L. The biology of thrombosis. *Annu. Rev. Med., 33*:479, 1982.

Olson, R. E. The function and metabolism of vitamin K. *Annu. Rev. Nutr., 4*:281, 1984.

Ratnoff, O. D. Thrombosis and the hypercoagulable state. *Circulation, 70*:III72, 1984.

Rosenberg, R. D. Role of heparin and heparinlike molecules in thrombosis and atherosclerosis. *Fed. Proc., 44*:A404, 1985.

Seegers, W. H. Blood clotting mechanisms: Three basic reactions. *Annu. Rev. Physiol., 31*:269, 1969.

Sigel, B., *et al.* Ultrasonic detection of red cell aggregation immediately preceding blood clotting. *Invest. Radiol., 19*:458, 1984.

Thomson, J. M. (ed.): Blood Coagulation and Haemostasis: A Practical Guide. London, Churchill Livingstone, 1979.

Wall, R. T., and Harker, L. A.: The endothelium and thrombosis. *Annu. Rev. Med., 31*:361, 1980.

III

MEMBRANE PHYSIOLOGY, NERVE, AND MUSCLE

9 ■ Transport Through the Cell Membrane

10 ■ Membrane Potentials and Action Potentials

11 ■ Contraction of Skeletal Muscle

12 ■ Neuromuscular Transmission; Function of Smooth Muscle

9

Transport Through the Cell Membrane

The fluid inside the cells of the body, called *intracellular fluid*, is very different from that outside the cells, called *extracellular fluid*. The extracellular fluid includes both the *interstitial fluid* that circulates in the spaces between the cells and also fluid of the *blood plasma* that mixes freely with the interstitial fluid through the capillary walls. It is the extracellular fluid that supplies the cells with nutrients and other substances needed for cellular function. But before the cell can utilize these substances, they must also be transported through the cell membrane.

Figure 9–1 gives the approximate compositions of both the extracellular and intracellular fluids. Note that the extracellular fluid contains large quantities of *sodium* but only small quantities of *potassium*. Exactly the opposite is true of the intracellular fluid. Also, the extracellular fluid contains large quantities of chloride, while the intracellular fluid contains very little. But the concentrations of phosphates, essentially all of which are organic metabolic intermediates, and proteins in the intracellular fluid are considerably greater than in the extracellular fluid. These differences between the components of the intracellular and extracellular fluids are extremely important to the life of the cell. It is the purpose of this chapter to explain how these differences are brought about by the transport mechanisms in the cell membrane.

The Lipid Barrier and the Transport Proteins of the Cell Membrane

The structure of the cell membrane was discussed in Chapter 2 and shown in Figure 2–3. Basically, it consists of a *lipid bilayer* with large numbers of protein molecules floating in the lipid, many if not most of them penetrating all the way through, as illustrated in Figure 9–2.

The lipid bilayer is not miscible with either the extracellular or the intracellular fluid. Therefore, it constitutes a barrier for the movement of most water-soluble substances between the extracellular and intracellular fluid compartments. However, as illustrated by the left-hand arrow of Figure 9–2, a few substances can penetrate this bilayer and can either enter the cell or leave it, passing directly through the lipid substance itself.

The protein molecules, on the other hand, have entirely different transport properties. Their molecular structures interrupt the continuity of the lipid bilayer and therefore constitute an alternate pathway through the cell membrane. Most of these proteins, therefore, are *transport proteins*. Different proteins function differently. Some have watery spaces all the way through the molecule and allow free movement of certain ions or molecules; these are called *channel proteins*. Others, called *carrier proteins*, bind with substances that are to be transported, and conformational changes in the protein molecules then move the substances through the interstices of the molecules to the other side of the membrane. Both the channel

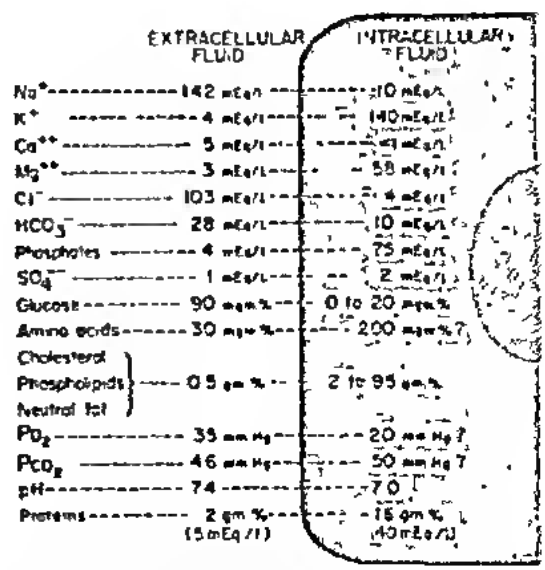


Figure 9–1. Chemical compositions of extracellular and intracellular fluids.

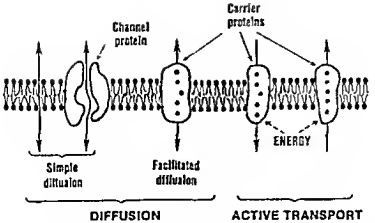


Figure 9–2. Transport pathways through the cell membrane, and the basic mechanisms of transport

proteins and the carrier proteins are highly selective in the type or types of molecules or ions that are allowed to cross the membrane.

Diffusion Versus Active Transport. Transport through the cell membrane, either directly through the lipid bilayer or through the proteins, occurs by one of two basic processes, *diffusion* (which is also called "passive transport") or *active transport*. Though there are many different variations of these two basic mechanisms, as we shall see later in this chapter, diffusion means random molecular movement of substances either through openings in the membrane or in combination with a carrier protein caused by the normal kinetic motion of matter. By contrast, active transport means movement of ions or other substances across the membrane in combination with a carrier protein but additionally *against an energy gradient*, such as from a low concentration state to a high concentration state, a process that requires chemical energy to cause the movement. Let us explain in more detail the basic physics and physical chemistry of these two separate processes.

DIFFUSION

All molecules and ions in the body fluids, including both water molecules and dissolved substances, are in constant motion, each particle moving its own separate way. Motion of these particles is what physicists call heat—the greater the motion, the higher the temperature—and motion never ceases under any conditions except at absolute zero temperature. When a moving molecule, A, approaches a stationary molecule, B, the electrostatic and internuclear forces of molecule A repel molecule B, adding some of the energy of motion to molecule B. Consequently, molecule B gains kinetic energy of motion while molecule A slows down, losing some of its kinetic energy. Thus, as shown in Figure 9–3, a single molecule in solution bounces among the other molecules first in one direction, then another, then another, and so forth, bouncing randomly billions of times each second.

This continual movement of molecules among each other in liquids, or in gases, is called *diffusion* Ions diffuse in exactly the same manner as whole molecules, and even suspended colloid particles diffuse in a similar manner, except that they diffuse far less rapidly than molecular substances because of their very large sizes.

DIFFUSION THROUGH THE CELL MEMBRANE

Simple Diffusion

Diffusion through the cell membrane is divided into two separate subprocesses called *simple diffusion* and *facilitated diffusion*. Simple diffusion means the molecular kinetic movement of molecules or ions through a membrane opening without the necessity of binding with carrier proteins in the membrane. The rate of diffusion is determined by the amount of substance available, by the velocity of kinetic motion, and by the number of openings in the cell membrane through which the molecules or ions can move. On the other hand, facilitated diffusion requires the interaction of the

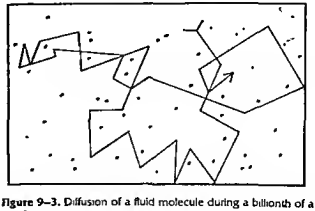


Figure 9–3. Diffusion of a fluid molecule during a billionth of a second.

molecules or ions with a carrier protein that aids its passage through the membrane, probably by binding chemically with it and shuttling it through the membrane in this form.

Simple diffusion can occur through the cell membrane by two pathways through the interstices of the lipid bilayer and through watery channels in some of the transport proteins, as illustrated to the left in *Figure 9–2.*

Diffusion Through the Lipid Bilayer

Diffusion of Lipid-Soluble Substances. In experimental studies, the lipids of cells have been separated from the proteins and then reconstituted as artificial membranes consisting of a lipid bilayer but without any transport proteins. Using such an artificial membrane, the transport properties of the lipid bilayer by itself have been determined.

One of the most important factors that determines how rapidly a substance will move through the lipid bilayer is the lipid solubility of the substance. For instance, the lipid solubilities of oxygen, nitrogen, and alcohols are very high, so that all these can dissolve directly in the lipid bilayer and diffuse through the cell membrane in exactly the same manner that diffusion occurs in a watery solution. For obvious reasons, the rate of diffusion of these substances through the membrane is directly proportional to their lipid solubility. Note especially that tremendous quantities of oxygen can be transported in this way; therefore, oxygen is delivered to the interior of the cell almost as though the cell membrane did not exist.

Transport of Water and Other Lipid-Insoluble Molecules. Even though water is highly insoluble in the membrane lipids, nevertheless it penetrates the cell membrane very readily, most of it passing directly though the lipid bilayer and still more passing through protein channels as well. The rapidity with which water molecules can penetrate the cell membrane is astounding. As an example, the total amount of water that diffuses in each direction through the red cell membrane during each second is approximately 100 times as great as the volume of the red cell itself.

The reason for the extreme diffusion of water through the lipid bilayer is still not certain, but it is believed that the water molecules are small enough and their kinetic energy great enough that they can simply penetrate like bullets though the lipid portion of the membrane before the "hydrophobic" character of the lipids can stop them.

Other lipid-insoluble molecules also can pass through the lipid bilayer in the same way as water molecules if they are small enough. However, as they become larger, their penetration falls off extremely rapidly. For instance, the diameter of the urea molecule is only 20 per cent greater than that of water. Yet its penetration of the cell membrane is about a thousand times less than that of water. Even so, remembering the astonishing rate of water penetration, this amount of penetration still allows rapid transport of urea through the cell membrane. The glucose molecule, which has a diameter of less than three times that of the water molecule, penetrates the lipid bilayer 100,000 times less rapidly than water, thus illustrating that the only lipid-insoluble molecules that can penetrate the lipid bilayer are the very small ones.

Failure of Ions to Diffuse Through the Lipid Bilayer. Even though water and other very small uncharged molecules diffuse easily through the lipid bilayer, ions—even small ones such as hydrogen ions, sodium ions, potassium ions, and so forth—penetrate the lipid bilayer about one million times less rapidly than does water. Therefore, any significant transport of these through the cell membrane must occur through channels in the proteins, as we shall discuss shortly.

The reason for the impenetrability of the lipid bilayer to ions is the electrical charge of the ions; this impedes ionic movement in two separate ways: (1) The electrical charge of these ions causes multiple molecules of water to become bonded to the ions, forming so-called *hydrated ions.* This greatly increases the sizes of ions, which alone impedes penetration of the lipid bilayer. (2) Even more important, the electrical charge of the ion also interacts with the charges of the lipid bilayer in the following way. It will be recalled that each half of the bilayer is composed of "polar" lipids that have an excess of negative charge facing toward the surfaces of the membrane and an excess of positive charge in the central portion of the membrane. Therefore, when a charged ion tries to penetrate either the negative or the positive electrical barrier, it is instantaneously repulsed.

To summarize, Table 9–1 gives the relative permeabilities of the lipid bilayer to a number of different molecules or ions of different diameters. Note especially the *extremely poor permeance of the ions* because of their electrical charges and the *poor permeance of glucose* because of its molecular diameter. But note also that glycerol penetrates the membrane almost as easily as urea even though its diameter is almost twice as great. The reason for this is a slight degree of lipid solubility.

Table 9–1 RELATIONSHIP OF EFFECTIVE DIAMETERS OF DIFFERENT SUBSTANCES TO THEIR LIPID BILAYER PERMEABILITIES

Substance	Diameter (nm)	Relative Permeability
Water molecule	0.3	1.0
Urea molecule	0.36	0.0006
Hydrated chloride ion	0.386	0.00000001
Hydrated potassium ion	0.396	0.0000000006
Hydrated sodium ion	0.512	0.0000000002
Glycerol	0.62	0.0006
Glucose	0.86	0.000009

Diffusion Through Protein Channels, and "Gating" of These Channels

The protein channels are believed to be watery pathways through the interstices of the protein molecules. Therefore, substances can diffuse directly through these channels from one side of the membrane to the other. However, the protein channels are distinguished by two important characteristics: (1) They are often selectively permeable to certain substances. (2) Many of the channels can be opened or closed by *gates*.

Selective Permeability of Different Protein Channels. Most, but not all, protein channels are highly selective for the transport of one or more specific ions or molecules. This results from the characteristics of the channel itself, such as its diameter, its shape, and the nature of the electrical charges along its surfaces. To give an example, one of the most important of the protein channels, the so-called *sodium channels*, are only 0.3 by 0.5 nm in size, but more importantly the inner surfaces of these channels are *strongly negatively charged*, as illustrated by the negative signs in the top panel of Figure 9–4. These strong negative charges pull the sodium ions more then they pull other physiologically important ions into the channels, because the ratio of pulling force to ionic diameter is far greater for sodium than for the others. Once in the channel, the sodium ion then diffuses in either direction according to the usual laws of diffusion. Thus, the sodium channel is specifically selective for the passage of sodium ions.

On the other hand, another set of protein channels is selective for potassium transport, illustrated in the lower panel of Figure 9–4. These channels are slightly smaller than the sodium channels, only 0.3 by 0.3 nm, but *they are not negatively charged*. Therefore, no strong attractive force is pulling ions into the channels. On the other hand, the hydrated form of the potassium ion is considerably smaller than the hydrated form of sodium because the sodium ion has one whole orbital set of electrons less than the potassium ion, which allows the sodium nucleus to attract far more water molecules than can the potassium. Therefore, the smaller hydrated potassium ions can pass easily through this channel whereas sodium ions are mainly rejected, thus once again providing selective permeability for a specific ion.

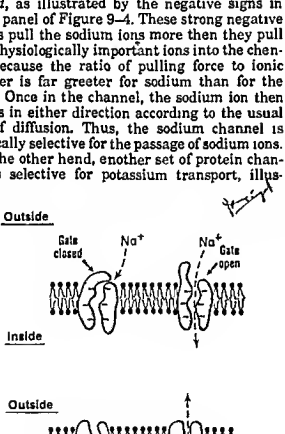


Figure 9–4. Transport of sodium and potassium ions through protein channels. Also shown are conformational changes of the channel protein molecules that open or close the "gates" guarding the channels.

Gating of Protein Channels. Gating of protein channels provides a means for controlling the permeability of the channels. This is illustrated in the upper and lower panels of Figure 9–4 for both the sodium and the potassium ion It is believed that the gates are actual gate-like extensions of the transport protein molecule, which can close over the opening of the channel or can be lifted away from the opening by a conformational change in the shape of the protein molecule itself. In the case of the sodium channels, this gate opens and closes on the outer surface of the cell membrane, whereas for the potassium channels it opens and closes on the inner surface

The opening and closing of gates are controlled by at least two principal ways:

(1) *Voltage gating.* In this instance, the molecular conformation of the gate responds to the electrical potential across the cell membrane. For instance, when there is a strong negative charge on the inside of the cell membrane, the sodium gates remain tightly closed; yet when the inside of the membrane loses its negative charge these gates open suddenly and allow tremendous quantities of sodium to pass inward through the sodium pores (until still another set of gates at the cytoplasmic ends of the channels closes them, as explained in Chapter 10). This is the basic cause of action potentials in nerves that are responsible for nerve signals. The potassium gates also open when the cell membrane becomes positively charged inside the cell. However, this response is different from that of the sodium gates for they open much more slowly; the resulting outflow of potassium ions from the cell helps in the recovery of the nerve fiber membrane at the end of the action potential. These events will be discussed in the following chapter.

(2) *Ligand gating.* Some protein channel gates are opened by the binding of another molecule with the protein, thus causing a conformational change in the protein molecule that opens or closes the gate. This is called *ligand gating*, substance that binds is the *ligand*. One of important instances of ligand gating is of acetylcholine on the so-called *acetylcholine*. This opens the gate of this channel,

a pore about 0 65 nm in diameter that allows all molecules and positive ions smaller than this diameter to pass through. This gate is exceedingly important in the transmission of signals from one nerve cell to another (Chapter 46) and from nerve cells to muscle cells (Chapter 12).

Figure 9–5 illustrates an especially interesting characteristic of voltage-gated channels. This figure shows two recordings of electrical current flowing through a sodium channel when there was an approximate 25 millivolt potential gradient across the pore. Note that the pore conducts current either all or none. That is, the gate of the pore snaps open and then snaps closed, each snapping event occurring within a few millionths of a second. This illustrates the rapidity with which the conformational changes can occur in the shape of the protein molecular gates. The figure also shows in the lower recording another important characteristic: at one voltage potential the pore may remain closed all the time or almost all the time, whereas at another voltage level it may remain open either all or most of the time. However, at in-between voltages, the pores tend to snap open and closed as illustrated in the upper recording, giving an average current flow somewhere between the minimum and the maximum.

Facilitated Diffusion

Facilitated diffusion is also called *carrier-mediated diffusion*, because a substance transported in this manner usually cannot pass through the membrane without a specific carrier protein helping it. That is, the carrier *facilitates* the diffusion of the substance to the other side.

Facilitated diffusion differs from simple diffusion through an open channel in the following very important way: whereas the rate of diffusion through an open channel increases proportionately

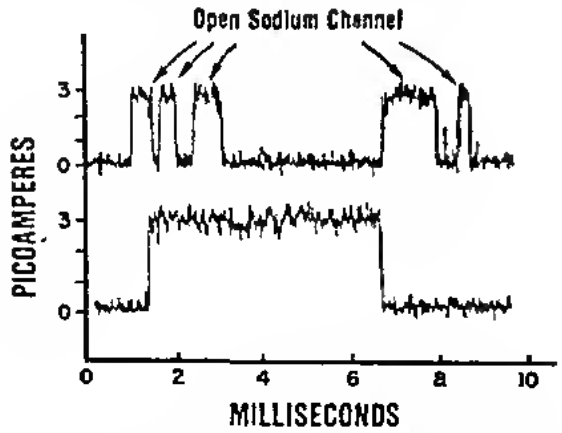


Figure 9–5. Record of current flow through a single voltage-gated sodium pore, demonstrating the all-or-none principle for opening the pore. (This type of record is obtained using a "patch-clamp," which is a minute pipette the tip of which abuts tightly against a small area of the cell membrane and allows current flow to be measured through individual pores.)

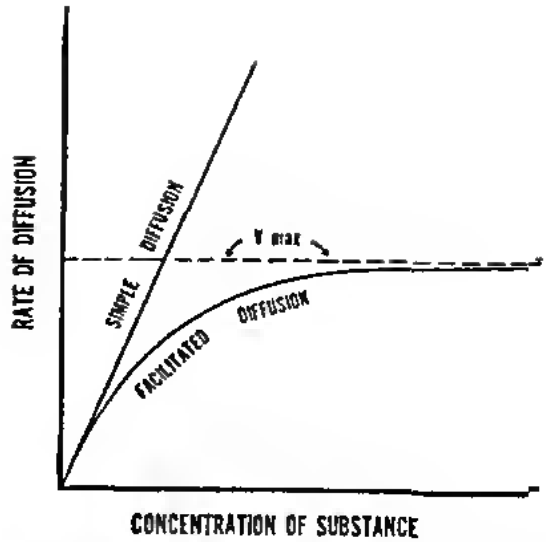


Figure 9–6. Effect of concentration of a substance on rate of diffusion through (a) a membrane in which there is simple diffusion and (b) a membrane in which there is facilitated diffusion. This illustrates that facilitated diffusion approaches a maximum rate called the V_{max}.

with the concentration of the diffusing substance, in facilitated diffusion the rate of diffusion approaches a maximum, called V_{max}, as the concentration of the substance increases. This difference between simple diffusion and facilitated diffusion is illustrated in Figure 9–6, showing that as the concentration of the substance increases the rate of simple diffusion continues to increase proportionately, but showing also the limitation of facilitated diffusion to the V_{max} level.

What is it that limits the rate of facilitated diffusion? A probable answer is the mechanism illustrated in Figure 9–7. This shows a carrier protein with a channel large enough to transport a specific molecule part way but not all the way through the membrane. It also shows a binding "receptor" on the protein carrier. The molecule to be transported enters the channel and then be-

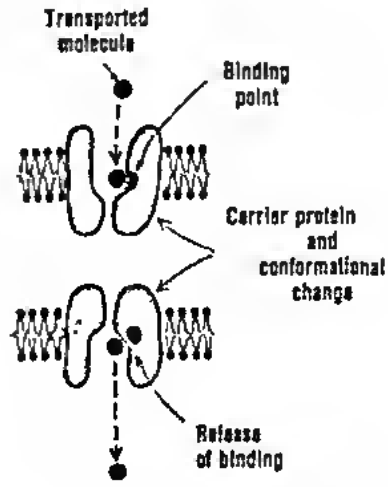


Figure 9–7. A postulated mechanism for facilitated diffusion.

comes bound. Then in a fraction of a second a conformational change occurs in the carrier protein so that the channel now opens to the opposite side of the membrane. Because the binding force of the receptor is rather weak, the thermal motion of the attached molecule causes it to break away and to be released on the opposite side. Obviously, the rate at which molecules can be transported by this mechanism can never be greater than the rate at which the carrier protein molecule can undergo conformational change back and forth between its two states; this is believed to occur at a constant maximum rate that is determined by the thermal motion of the carrier molecule. Note specifically that this mechanism allows the transported molecule to "diffuse" in either direction through the membrane.

Among the most important substances that cross cell membranes by facilitated diffusion are *glucose* and the *amino acids*. In the case of glucose, the carrier molecule is known to have a molecular weight of about 45,000; it can also transport several other monosaccharides that have structures similar to that of glucose, including mannose, galactose, xylose, and arabinose. Also, insulin can increase the rate of facilitated diffusion of glucose as much as 10- to 20-fold. This is the principal mechanism by which insulin controls glucose utilization in the body, as we shall discuss in Chapter 78.

NET DIFFUSION THROUGH THE PROTEIN CHANNELS OF THE CELL MEMBRANE AND FACTORS THAT AFFECT IT

By now it is evident that many different substances can diffuse either through the lipid bilayer of the cell membrane or through protein channels. However, please understand clearly that substances that diffuse in one direction can also diffuse in the opposite direction. Usually, what is important to the cell is not the total substance diffusing in both directions but the difference between these two, which is the *net rate of diffusion* in one direction. The factors that affect this are (1) the permeability of the membrane, (2) the difference in concentration of the diffusing substance between the two sides of the membrane, (3) the pressure difference across the membrane, and (4) in the case of ions, the electrical potential difference between the two sides of the membrane.

The Permeability of the Membrane. The permeability of a membrane is defined as the rate of transport through a unit area of membrane for a given concentration difference. Among the factors that affect the total permeability of the membrane protein channels are: (a) the *number* of channels through which the substance can diffuse, (b) the *length* of the channels, (c) the degree of impediment to movement through each channel—that is, the *resistance* of the channel, (d) the *molecular weight* of the diffusing substance, and (e) the *temperature*. The effect of each of these on the permeability is obvious, with the possible exception of temperature and molecular weight.

A higher temperature increases the permeability because the thermal motion of all the molecules and ions in a solution increases directly with the temperature, thus more diffusion through the channels.

The molecular weight affects diffusion of molecules and ions mainly because the velocity of kinetic motion in a solution is inversely proportional to the square root of the molecular weight. Therefore, in general, the permeability of the membrane is also approximately inversely proportional to the square root of the molecular weight.

The following formula gives the approximate effect of all the above factors together on the permeability of the membrane for diffusion of any given substance:

$$\text{Permeability} = \frac{\text{Number of channels per unit area} \times \text{temperature}}{\text{Resistance of channels per unit length} \times \text{length of channel} \times \text{square root of molecular weight}}$$

Effect of a Concentration Difference. Figure 9–8*A* illustrates a membrane with a substance in high concentration on the outside and low concentration on the inside. The rate at which the substance diffuses *inward* is proportional to the con-

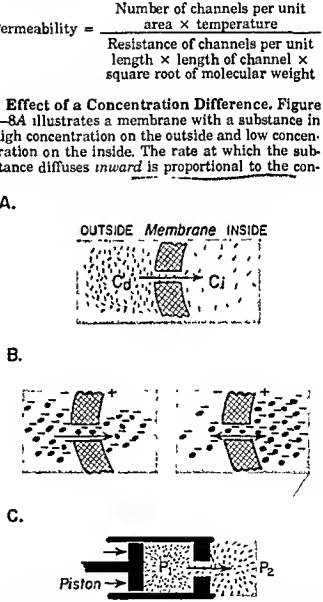


Figure 9–8. Effect of (*A*) concentration difference, (*B*) electrical difference, and (*C*) pressure difference on net diffusion of molecules and ions through a cell membrane.

centration of molecules on the outside, for this concentration determines how many of the molecules strike the outside of the channel each second. On the other hand, the rate at which the molecules diffuse *outward* is proportional to their concentration *inside* the membrane. Obviously, therefore, the rate of net diffusion into the cell is proportional to the concentration on the outside *minus* the concentratin on the inside or

$$\text{Net diffusion} \propto P\,(C_o - C_i)$$

in which C_o is the concentration on the outside, C_i is the concentration on the inside, and P is the permeability of the membrane for the substance.

Effect of an Electrical Potential Difference. If an electrical potential is applied across the membrane as shown in Figure 9–8*B*, because of their electrical charges ions will move through the membrane even though no concentration difference exists to cause their movement. Thus, to the left in tha figure, the concentrations of negative ions are exactly the same on both sides of the membrane, but a positive charge has been applied to the right side of the membrane and a negative charge to the left, creating an electrical gradient across tha membrane. The positive charge attracts the negative ions while the negative charge repels them Therefore, net diffusion occurs from left to right. After much time large quantities of negative ions will have moved to the right (if we neglect, for the time being, the disturbing effects of the positiva ions of the solution), creoting the condition illustrated on the right in Figure 9–8*B*, in which a concentration difference of the same ions has developed in the direction opposite to the electrical potential difference. Obviously, the concentration difference is now tending to move the ions to the left, while the electrical difference is tending to move them to the right When the concentration difference rises high enough, the two effects will exactly balance each other. At normal body temperature (38°C), the electrical differenre that will exactly balance a given concentration difference of *univalent* ions, such as Na^+, K^+, or Cl^-, can be determined from the following formula called the *Nernst equation.*

$$\text{EMF (in millivolts)} = \pm 61 \log \frac{C_1}{C_2}$$

in which EMF is the electromotive force (voltage) between side 1 and side 2 of the membrane, C_1 is the concentration on side 1, and C_2 is the conrentration on side 2 The polarity of the voltage on side 1 in the above equation is + for negative ions and − for positive ions. This relationship is extremely important in understanding the transmission of nerve impulses, for which reason it is discussed in even greater detail in Chapter 10.

Effect of a Pressure Difference. At times considerable pressure difference develops between the two sides of a membrane. This occurs, for instance, at the capillary membrane, which has a pressure approximately 23 mm Hg greater inside the capillary than outside. Pressure actually means the sum of all the forces of the different molecules striking a unit surface area at a given instant. Therefore, when the pressure is higher on one side of a membrane than the other, this means that the aum of all the forces of the molecules striking the channels on that side of the membrane is greater than on the other side. This can result either from greater numbers of molecules striking the membrane per second or from greater kinetic energy of the average molecule striking the membrane. In either event, increased amounts of energy are available to cause net movement of molecules from the high pressure side toward the low pressure side. This effect is illustrated in Figure 9–8C, which shows a piston developing high pressure on one side of a cell membrane, thereby causing net diffusion through the membrane to the other side.

For the usual red cell membrane, 1 mm Hg pressure difference causes approximately 10^{-4} cubic micron of net diffusion of water through each square micron of membrane each second. This appeors to be only a very minuta rate of fluid movement through the membrane, but in relation to the normal cell size and the very large diffusion pressure that can develop at the cell membrane because of osmotic forces, as discussed later in the chapter, this rote can represent tremendous transport of fluid in only a few seconds.

OSMOSIS ACROSS SELECTIVELY PERMEABLE MEMBRANES— NET DIFFUSION OF WATER

By far the most abundant substance to diffuse through the cell membrane is water. It should be recalled again that enough water ordinarily diffuses in each direction through the red cell membrane per second to equal about *100 times the volume of the cell itself* Yet, *normally*, the amount that diffuses in the two directions is so precisely balanced that not even the slightest *net* movement of water occurs. Therefore, the volume of the rell remains constant However, under certain conditions, a *concentration difference for water* can develop across a membrane, just as concentration differences for other substances can also occur. When this happens, net movement of water does occur across the cell membrane, causing the cell either to swell or to shrink, depending on the direction of the net movement. This process of net movement of water caused by a concentration difference of water is called *osmosis.*

To give an example of osmosis, let us assum the conditions shown in Figure 9–9, with pur water on one side of the cell membrane and a solution of sodium chloride on the other side Referring back to Table 9–1, we see that wate

tration in terms of mass is of no value in determining osmotic pressure To express the concentration in terms of numbers of particles, the unit called the *osmole* is used in place of grams.

One osmole is the number of particles (molecules) in 1 gram molecular weight of undissociated solute Thus, 180 grams of glucose, which is one gram molecular weight of glucose, is equal to 1 osmole of glucose, because glucose does not dissociate On the other hand, if the solute dissociates into two ions, 1 gram molecular weight of the solute equals 2 osmoles, because the number of osmotically active particles is now twice as great as is the case in the undissociated solute. Therefore, 1 gram molecular weight of sodium chloride, 58.5 grams, is equal to 2 osmoles.

A solution that has *1 osmole of solute dissolved in each kilogram of water* is said to have an *osmolality of 1 osmole per kilogram*, and a solution that has 1/1000 osmole dissolved per kilogram has an osmolality of 1 milliosmol per kilogram The normal osmolality of the extracellular and intracellular fluids is about 300 milliosmoles per kilogram.

Relationship of Osmolality to Osmotic Pressure. At normal body temperature, 38° C, a concentration of 1 osmole per liter will cause *19,300* mm *Hg* osmotic pressure in the solution Likewise, 1 milliosmole per liter concentration is equivalent to *19.3* mm *Hg* osmotic pressure Multiplying this value times the 300 milliosmolar concentration of the body fluids gives a total calculated osmotic pressure of these fluids of 5790 mm Hg. The measured values for this, however, average only about 5500 mm Hg The reason for this difference is that many of the ions in the body fluids, such as the sodium and the chloride ions, are highly attracted to each other; consequently, they cannot move totally unrestrained in the fluids and create their full osmotic potential Therefore, on the average, the actual osmotic pressure of the body fluids is about 0 93 times the calculated value

The Term "Osmolarity." Because of the difficulty of measuring kilograms of water in a solution, which is required to measure the "osmolality," one usually uses another term, "osmolarity," which is the osmolar concentration expressed as *osmoles per liter of solution* rather than osmoles per kilogram of water Though, strictly speaking, it is osmoles per kilogram of water (osmolality) that determines the osmotic pressure, nevertheless, for dilute solutions such as occur in the body, the quantitative differences between osmolarity and osmolality are less than 1 per cent Since it is far more practical to measure osmolarity than osmolality, this is the usual practice in almost all physiological studies

A Quandary: How Is It Possible for the Osmotic Pressure of a Solution to Be Greater Than the Atmospheric Pressure? The discerning student will have noted that the osmotic pressure of the body fluids is greater than the atmospheric pressure If this is true, why is it that the body does not explode?

The answer to this question is that osmotic pressure is not the force of water molecules attempting to escape out of the body, because the water molecules are mutually attracted to each other, which mostly nullifies their escape force except for the so-called "vapor pressure" of the water (see Chapter 40). On the other hand the mutual attraction does not diminish the force of the water molecules attempting to pass through water channels in a membrane because the molecules are also attracted by the molecules inside the channels The simple analysis in Figure 9–11 can be helpful in understanding this difference.

The dashed line in the figure represents a membrane selectively permeable to water molecules; to the left represented a very large molecule that cannot pass through the membrane, and in the center are small water molecules that can go through The presence of large molecules above the membrane has diluted the water molecules so that the concentration of water molecules above the membrane is less than the concentration below Putting numbers to this, the pure water beneath the membrane has a concentration of 55 osmoles per liter Multiplying this by 19,300 gives a total *potential pressure* of the water molecules beneath the membrane of 1,073,000 mm Hg, which is a measure of the so-called "chemical potential" of the water. Now for the top side of the membrane· In the body fluids there are approximately 0 3 osmole of dissolved substances, as we just calculated this accounts for 5500 mm Hg of potential pressure Since this dilutes the water molecules by this same amount, subtracting 5500 from 1,073,000 mm Hg leaves a *potential pressure* of the diluted water molecules above the membrane of only 1,067,500 mm Hg Thus, the forces of the water molecules striking the membrane both from below and from above are tremendous, but the difference between these two forces is only a small amount and is equal to the force exerted by the nondiffusible particles. Therefore to the right in Figure 9–11, it is shown that the *pressure difference* across the membrane that is forcing water molecules through the water channels is equal to 5500 mm Hg This is called the *osmotic pressure*, but it is actually a pressure difference that is causing net diffusion of water molecules in one direction through the membrane

"Bulk Flow" of Water Through Membrane Channels in Response to Hydrostatic and Osmotic Pressure Gradients

When either a hydrostatic or an osmotic pressure gradient is created across a membrane, the rate of movement of water molecules through the membrane

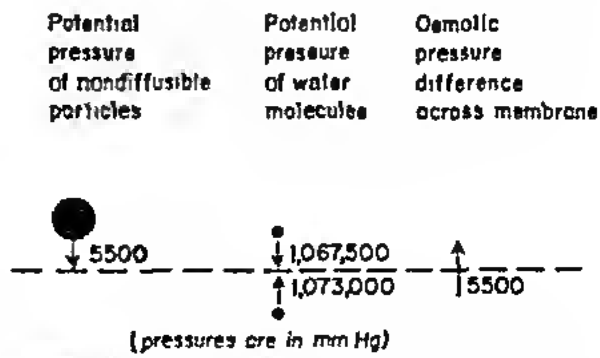


Figure 9–11. Kinetics of osmosis. (See text for discussion)

channels is often many times as great as the rate that can be accounted for by net diffusion alone. This is caused by a phenomenon called "bulk flow," which means simply that since large numbers of molecules are moving in the same direction, they tend to "stream" through the pores in unison rather than to move randomly as occurs in pure diffusion. As examples of this phenomenon, the osmotic movement of water through the red cell roembrane is about four times as great and through the capillary membrane as much as ten times as great as can be accounted for by simple diffusion. Bulk flow will be discussed in greater detail in Chapter 30 in relation to the dynamics of transport through the capillary membrane.

ACTIVE TRANSPORT

Often only a minute concentration of a substance is present in the extracellular fluid, and yet a large concentration of the substance is required in the intracellular fluid. For instance, this is truo of potassium ions. Conversely, other substances frequently enter cells and must bo removed even though their concentrations inside are far less than outside. This is truo of sodium ions.

From the discussion thus far it is evident that *no substances con diffuse ogoinst an "electrochemical gradient,"* which is the sum of all the diffusion forces acting at the membrane—the forces caused by concentration difference, electrical difference, and pressura difference. That is, it is often said that substances cannot diffuse "uphill." To cause movement of a substance uphill, energy must be imparted to tha substance. This is analogous to tho compression of air by a pump. Compression causes the concentration of tho air molecules to increase but to creato this greater concentration, anergy must be imparted to the air molecules by the piston of the pump as they are compressed. Likewise, as molecules are transported through a cell membrane from a dilute solution to a concentrated solution, energy must be imparted to the molecules to concentrate them. When a cell membrane moves molecules uphill against a concentration gradient (or uphill against an electrical or pressure gradient) the process is called *active transport.*

Among the different substances that are actively transported through cell membranes are sodium ions, potassium ions, calcium ions, iron ions, hydrogen ions, chloride ions, iodide ions, urate ions, severol different sugors, and the amino acids.

BASIC MECHANISM OF ACTIVE TRANSPORT

Active transport depends on transport of substances by *carrier proteins* that penetrate through the membrane, the same as is true for facilitated diffusion. However, in active transport, the carrier functions differently from the carrier in facilitated diffusion, because the carrier protein must now impart energy to the substance that is being transported to move it against an electrochemical gradient. The energy for this purpose is derived from ATP. Furthermore, the active transport carrier proteins have ATPase activity, which means that they can cleave ATP to form ADP or AMP with release of the energy from the high energy phosphate bonds. Unfortunately, though, the manner in which this energy is coupled to cause transport of the substance against the electrochemical gradient is not yet clear.

The Sodium-Potassium "Pump"

The most ubiquitous active transport mechonism in the body is the one that transports sodium ions out of cells to the exterior and at the same time pumps potassium ions from the outside to the inside; it is called the *sodium-potassium pump.* This pump is present in oil cells of the body, and it is responsible for maintaining the sodium and potassium concentration differences across the cell membrane as well as for establishing a negative electrical potential inside the cells. Indeed, we shall seo in tho next chapter that this pump is the basis of nerve function to transmit signals throughout the nervous system.

Figure 9–12 illustrates the basic components of the Na^+-K^+ pump. The *carrier protein* is a complex of two separate globular proteins, a larger ona with a molecular weight of about 100,000 and a smaller one with a molecular weight of 45,000. Though the function of the smaller protein is not known, the lorger protein has three specific features thot are importont for function of the pump:

(1) It has three *receptor sites for binding sodium ions* on the portion of tha protein that protrudes to the interior of the cell.

(2) It has two *receptor sites for potassium ions* on the outside.

(3) The inside portion of this protein adjacent to or near to the sodium binding sites has ATPase activity.

Now to put the pump into perspective: When three sodium ions bind on the inside of the carrier

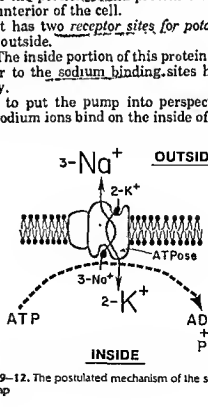


Figure 9–12. The postulated mechanism of the sodium-potassium pump

protein and two potassium ions on the outside, the ATPase function of the protein becomes activated. This then cleaves 1 molecule of ATP, splitting it to ADP and liberating a high energy phosphate bond of energy. This energy is then believed to cause a conformational change in the protein carrier molecule, extruding the sodium ions to the outside and the potassium ions to the inside.

Note also in Figure 9–12 that the concentration of sodium on the inside, as designated by the small size of the Na^+ symbol, is very slight, whereas the concentration of the sodium on the outside is very great. Conversely, the potassium ion concentration on the outside is very slight whereas its concentration on the inside is very great. Thus, both these ions are transported against very large energy gradients.

Importance of the Na^+-K^+ Pump in Controlling Cell Volume. One of the most important functions of the sodium-potassium pump is to control the volume of the cells. Without function of this pump, most cells of the body would swell until they burst. The mechanism for controlling the volume is the following: Inside the cell are large numbers of proteins and other organic compounds that cannot escape from the cell. Most of these are negatively charged and therefore collect around them large numbers of positive ions as well. All these substances then tend to cause osmosis of water to the interior of the cell, and unless this is checked, the cell will swell indefinitely until it bursts. However, the normal mechanism for preventing this is the Na^+-K^+ pump. Note again that this pump pumps 3 Na^+ ions to the outside of the cell for every 2 K^+ ions pumped to the interior. Also, the membrane is far less permeable to sodium ions than to potassium ions, so that once the sodium ions are on the outside they have a strong tendency to stay there. Thus, this represents a continual net loss of ionic substances out of the cell, which initiates an opposite osmotic tendency to move water out of the cell. Furthermore, when the cell begins to swell, this automatically activates the sodium-potassium pump, moving still more ions to the exterior and carrying water with them. Therefore, the sodium-potassium pump performs a continual surveillance role in maintaining normal cell volume.

The Electrogenic Nature of the Na^+-K^+ Pump. The fact that the Na^+-K^+ pump moves three sodium ions to the exterior for every two potassium ions to the interior means that a net of one positive charge is moved from the interior of the cell to the exterior for each revolution of the pump. This obviously creates positivity outside the cell but leaves a deficit of positive ions inside the cell, that is, it causes negativity on the inside. Therefore, the sodium-potassium pump is said to be *electrogenic*, because it creates an electrical potential across the cell membrane as it pumps.

The Calcium Pump

Another very important pump is the calcium pump. Calcium ions are normally maintained at extremely low concentration in the intracellular fluid, at a concentration about 10,000 times less than that in the extracellular fluid. This is achieved by two calcium pumps: One is in the cell membrane and pumps calcium to the outside of the cell. The other pumps calcium ions into one or more of the internal vesicular organelles of the cell, such as into the sarcoplasmic reticulum of muscle cells and into the mitochondria in all cells. In both instances, the carrier protein responsible for pumping the calcium ions is an ATPase having the same capability to cleave ATP as the ATPase sodium carrier protein. The difference is that this protein has a binding site for calcium instead of sodium.

Saturation of Active Transport

Active transport saturates in exactly the same way that facilitated diffusion saturates. When the concentration of the substance to be transported is small, the rate of transport increases approximately in proportion to the increase in concentration. However, at large concentrations, transport approaches a maximum, which is also called the V_{max}, as is also true for facilitated diffusion. The saturation is caused by limitation of the rates at which the chemical reactions of binding, release, and carrier conformational changes can occur.

Energetics of Active Transport

The amount of energy required to transport a substance actively through a membrane (aside from energy lost as heat in the chemical reactions) is determined by the degree that the substance is concentrated during transport. Compared to the energy required to concentrate a substance tenfold, to concentrate it 100-fold requires twice as much energy, and to concentrate it 1000-fold requires three times as much. In other words, the energy required is proportional to the logarithm of the degree that the substance is concentrated, as expressed by the following formula:

$$\text{Energy (in calories per osmole)} = 1400 \log \frac{C_1}{C_2}$$

That is, in terms of calories, the amount of energy required to concentrate 1 osmole of substance tenfold is about 1400 calories. One can see that the energy expenditure for concentrating substances in cells or for removing substances from cells against a concentration gradient can be tremendous. Some cells, such as those lining the renal tubules as well as many glandular cells, perhaps expend as much as 90 per cent of their energy for this purpose alone.

SECONDARY ACTIVE TRANSPORT: SODIUM "COTRANSPORT" OF GLUCOSE AND AMINO ACIDS

Glucose and amino acids are transported into some cells against very large concentration gradients, even though the carrier protein for transport of these substances cannot cleave ATP to provide the required energy. Instead, the energy is provided by the concentration gradient of the sodium ions across the cell membrane. The mechanism of this is illustrated in Figure 9–13. Note that the carrier protein has two binding sites on its exterior side, one for sodium and one for glucose. But note also that the concentration of sodium is very high on the outside and very low on the inside. Therefore, sodium ions are always attempting to move down their energy gradient to the interior of the cell, causing continual stress on the carrier protein to let the sodium ions pass on through to the interior. However, another property of this carrier protein is that the conformational change of the protein molecule will not occur to allow the sodium to move to the interior unless a glucose molecule is also attached. But once both the sodium ion and the glucose molecule are attached, then the conformational change takes place and both the sodium and glucose are transported to the inside of the cell at the same time. Furthermore, this can occur even though the glucose must be transported "uphill" from a low concentration on the outside to a high concentration inside. That is, the energy gradient of the sodium ions literally pulls the glucose to the interior, thus utilizing energy from the sodium gradient to cause the uphill transport of the glucose. For obvious reasons, this type of transport has received both the names *secondary active transport* and *sodium-glucose cotransport*.

Sodium cotransport of the amino acids occurs in the same manner as for glucose except that it utilizes a different set of transport proteins. Five separate *amino acid transport proteins* have been identified, each of which is responsible for transporting one subset of amino acids with specific molecular characteristics.

Sodium cotransport of glucose and amino acids occurs especially in the epithelial cells of the intestinal tract and renal tubules to aid in the

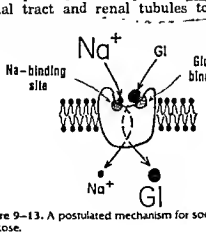


Figure 9–13. A postulated mechanism for sodium cotransport of glucose.

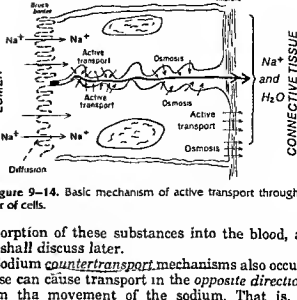


Figure 9–14. Basic mechanism of active transport through a layer of cells.

absorption of these substances into the blood, as we shall discuss later.

Sodium *countertransport* mechanisms also occur; these can cause transport in the *opposite direction* from the movement of the sodium. That is, a receptor site is present on the outside of the carrier protein for sodium but on the inside for the other substance, and movement of the sodium down its energy gradiant to the inside of the cell causes the other substance to be transported to the exterior. Transport of this type has been observed in some cells for exchange of sodium and other ions, such as potassium or calcium, across the membrane.

It is clear that the ultimate source of the energy for sodium cotransport or countertransport is once again ATP, which drives the sodium-potassium pump to build up the necessary sodium concentration gradient.

ACTIVE TRANSPORT THROUGH CELLULAR SHEETS

At many places in the body, substances must be transported through a cellular sheet instead of simply through the cell membrane. Transport of this type occurs through the intestinal epithelium, the epithelium of the renal tubules, the epithelium of all exocrine glands, the epithelium of the gall bladder, the membrane of the choroid plexus of the brain, and many other membranes.

The basic mechanism for transport of a substance through a cellular sheet is (1) to provide *active transport* through the cell membrane *on one side* of the sheet and then (2) to provide for either *simple diffusion* or *facilitated diffusion* through the membrane on *the opposite side* of the sheet.

Figure 9–14 illustrates the mechanism for transport of sodium ions through the epithelium of the intestines, the gall bladder, and the renal tubules. This figure shows that the epithelial cells are connected together only at the luminal pole, but the junctions at this point are all tight junctions, thus mainly blocking sodium ion diffusion between the cells. However, the luminal surfaces of the cells are very permeable to both sodium ions and to water. Therefore, both sodium ions and water diffuse readily to the interior of the cell. Then, at the basal and lateral membranes of the cells the

sodium ions are actively transported into the extracellular fluid. This creates a high sodium ion concentration gradient across these membranes, which in turn causes osmosis of water as well. Thus, the active transport of sodium ions at the basolateral poles of the epithelial cells results not only in the transport of sodium ion but of water also.

Furthermore, any other substances that might be linked with the sodium can be transported in this same way. For instance, the positive charges of the sodium ions usually pull negatively charged chloride ions along with the sodium. Likewise, when glucose (or amino acids) is cotransported along with sodium through the luminal surface of the cell, the concentration of the glucose increases inside the cell. Then this glucose is transported by facilitated diffusion through the basaolateral borders of the cell, finally entering the extracellular fluid along with the sodium ions, the chloride ions, and the water.

Thus, these are the mechanisms by which almost all the nutrients, the ions, and other substances are absorbed into the blood from the intestine; they are also the way in which these same substances are reabsorbed from the glomerular filtrate by the renal tubules.

REFERENCES

Agnew, W S Voltage-regulated sodium channel molecules *Annu Rev Physiol*, *46* 517, 1984

Almers, W, and Stirling, C Distribution of transport proteins over animal cell membranes *J Membr Biol* *77* 169, 1984

Andersen, O S Gramicidin channels *Annu Rev Physiol*, *46* 531 1984

Andreoli, T E, *et al* (eds) Membrane Physiology New York Plenum Press, 1980

Armstrong, C M Sodium channels and gating currents *Physiol Rev*, *61* 644, 1981

Auerbach, A., and Sachs F Patch clamp studies of single ionic channels *Annu Rev Biophys Bioeng*, *13* 269, 1984

Barry, P H and Diamond, J M Effects of unstirred layers in membrane phenomena *Physiol Rev*, *64* 763, 1984

Blachley, J E, *et al* The harmful effects of ethanol on ion transport and cellular respiration *Am J Med Sci* *289* 22, 1985

Blaustein, M P, and Hamlyn, J M Sodium transport inhibition, cell calcium and hypertension The natriuretic hormone Na^+–Ca^{2+} exchange hypertension hypothesis *Am J Med*, *77* 45, 1984

Chase, H S, Jr Does calcium couple the apical and basolateral membrane permeabilities in epithelia? *Am J Physiol*, *247* F869, 1984

Dawson, K G Endocrine physiology of electrolyte metabolism *Drugs*, *28*(Suppl 1) 98 1984

DiPolo, R, and Beaugé, L The calcium pump and sodium-calcium exchange in squid axons *Annu Rev Physiol* 45 313, 1983

Ellis, D Na-Ca exchange in cardiac tissues *Adv Myocardiol*, *5* 295 1985

Ellory C, and Lew V L (eds) Membrane Transport in Red Cells. New York, Academic Press, 1977

Fettiplace, R and Haydon D A Water permeability of lipid membranes *Physiol Rev* *60* 510, 1980

her, K A Split membrane analysis. *Annu Rev Physiol*, *42* 261, 980

tman, P W Magnesium transport across cell membranes. *J Membr hol* *80* 1 1984

zell R A and Heintze K Transport functions of the gallbladder n Javitt, N B (ed) International Review of Physiology Liver and iliary Tract Physiology I Vol 21 Baltimore University Park Press, 980 p 221

zell R A, and Schultz, S G Models of electrolyte absorption and ecretion by gastrointestinal epithelia *In* Crane, R K (ed) International Review of Physiology Gastrointestinal Physiology III Vol 19 Baltimore, University Park Press, 1979, p 205

Gadsby, D C The Na/K pump of cardiac cells *Annu Rev. Biophys Bioeng*, *13* 373, 1984

Gilles, R (ed) Mechanisms of Osmoregulation. Maintenance of Cell Volume New York, John Wiley & Sons 1979

Goldin, S M, *et al*. Isolation and reconstitution of neuronal ion transport proteins *Annu Rev Neurosci*, *6* 419, 1983

Goresky, C A Uptake in the liver the nature of the process *In* Javitt, N B (ed) International Review of Physiology Liver and Biliary Tract Physiology I Vol 21 Baltimore, University Park Press, 1980, p 65

Gregor, H P, and Gregor, C D Synthetic membrane technology *Sci Am*, *239*(1) 112, 1978

Gunn, R B Co- and counter-transport mechanisms in cell membranes *Annu Rev Physiol*, *42* 249 1980

Guyton, A C, *et al* Dynamics and Control of the Body Fluids Philadelphia W B Saunders Co, 1975

Hess, G P, *et al* Acetylcholine receptor-controlled ion translocation. Chemical kinetic investigations of the mechanism *Annu. Rev Biophys Bioeng*, *12* 443, 1983

Hinz, H J Thermodynamics of protein ligand interactions Calorimetric approaches *Annu Rev Biophys Bioeng*, *12* 285, 1983

Hobbs, A S and Albers R W The structure of proteins involved in active membrane transport *Annu Rev Biophys Bioeng*, *9* 259, 1980

Ives, H E, and Rector, F C, Jr Proton transport and cell function *J Clin Invest*, *73* 285, 1984

Jacob, H S (ed) Blood Cell Membranes New York, Grune & Stratton, 1979

Katz, A M Basic cellular mechanisms of action of the calcium-channel blockers *Am J Cardiol*, *55* 2B, 1985

Keynes, R D Ion channels in the nerve-cell membrane *Sci Am*, *240*(3) 126, 1979

Kregenow, F M Osmoregulatory salt transporting mechanisms Control of cell volume in anisotonic media *Annu Rev Physiol*, *43* 493 1981

Latorre R and Alvarez, O Voltage-dependent channels in planar lipid bilayer membranes *Physiol Rev*, *61* 77, 1981

Lerner, J Cell membrane amino acid transport processes in the domestic fowl (*Gallus domesticus*) *Comp Biochem Physiol*, *78* 205, 1984

Lindermann, B Fluctuation analysis of sodium channels in epithelia. *Annu Rev Physiol*, *46* 497 1984

Lodish, H F and Rothman, J E The assembly of cell membranes *Sci Am* *240*(1) 48 1979

Macey, R I Transport of water and urea in red blood cells *Am J Physiol*, *246* C195 1984

Macknight, A D C and Leaf, A Regulation of cellular volume *Physiol Rev*, *57* 510, 1977

Malhotra, S K The Plasma Membrane New York, John Wiley & Sons, 1983

Miller C Integral membrane channels. Studies in model membranes *Physiol Rev*, *63* 1209, 1983

Miller, C Ion channels in liposomes *Annu Rev Physiol*, *46* 549, 1984

Naftalin, R J The thermostatics and thermodynamics of cotransport. *Biochem Biophys Acta*, *779* 155, 1984

Parker, J C and Berkowitz L R., Physiologically instructive genetic variants involving the human red cell membrane *Physiol Rev*, 63 261, 1983

Reuter, H Ion channels in cardiac cell membranes *Annu Rev. Physiol*, *46* 473, 1984

Sakmann, B, and Neher, E Patch clamp techniques for studying ionic channels in excitable membranes. *Annu Rev Physiol*, *46* 455, 1984

Schatzmann, H J The red cell calcium pump *Annu Rev Physiol*, *45* 303, 1983

Schultz, S G Principles of electrophysiology and their application to epithelial tissues *Int Rev Physiol*, *4* 69, 1974

Schultz S G A cellular model for active sodium absorption by mammalian colon *Annu Rev Physiol*, *46* 435, 1984

Schultz, S G, and Curran, P F Coupled transport of sodium and organic solutes *Physiol Rev*, *50* 637, 1970

Sjodin, R A Ion transport in skeletal muscle New York, John Wiley & Sons, 1982

Smith P L, and McCabe, R D, Mechanism and regulation of transcellular potassium transport by the colon. *Am J Physiol* *247* G445, 1984

Stekhoven, F. S, and Bonting, S L Transport adenosine triphosphatases Properties and functions. *Physiol Rev*, *61* 1, 1981

Ullrich, K J Sugar, amino acid, and Na^+ cotransport in the proximal tubule *Annu Rev Physiol*, *41* 181, 1979

Vermeer, C The vitamin K-dependent carboxylation reaction *Mol Cell Biochem*, *61* 17, 1984

Wright, E M Electrophysiology of plasma membrane vesicles *Am J Physiol*, *246* F363, 1984

Wright, E M, and Diamond, J M Anion selectivity in biological systems *Physiol Rev*, *57* 109, 1977

10

Membrane Potentials and Action Potentials

Electrical potentials exist across the membranes of essentially all cells of the body, and some cells, such as nerve and muscle cells, are "excitable"—that is, capable of self-generation of electrochemical impulses at their membranes and, in some instances, employment of these impulses to transmit signals along the membranes In still other types of cells, such as glandular cells, macrophages, and ciliated cells, changes in membrane potentials probably play significant roles in controlling many of the cell's functions. However, the present discussion is concerned with membrane potentials generated both at rest and during action by nerve and muscle cells.

BASIC PHYSICS OF MEMBRANE POTENTIALS

Before beginning this discussion, let us first recall that the fluids both inside and outside the cells are electrolytic solutions containing 150 to 160 mEq per liter of positive ions and the same concentration of negative ions. Generally, a very minute excess of negative ions (anions) accumulates immediately inside the cell membrane along its inner surface, as illustrated in Figures 10–1A and 10–2, and an equal number of positive ions (cations) accumulates immediately outside the membrane. The effect of this is the establishment of a *membrane potential* between the inside and outside of the cell.

The two basic means by which membrane potentials can develop are (1) diffusion of ions through the membrane as a result of ion concentration differences between the two sides of the membrane, thus creating an imbalance of negative and positive charges on the two sides of the membrane, and (2) active transport of ions through the membrane, thus also creating an imbalance of charges.

MEMBRANE POTENTIALS CAUSED BY DIFFUSION

Figures 10–1A and B illustrate a nerve fiber when there is no active transport of either sodium

DIFFUSION POTENTIALS

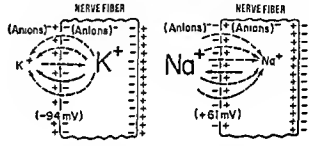


Figure 10–1. *A,* Establishment of a diffusion potential across a cell membrane, caused by potassium ions diffusing from inside the cell to the outside through a membrane that is selectively permeable only to potassium *B* Establishment of a diffusion potential when the membrane is permeable only to sodium ions *Note that the internal membrane potential is negative when* potassium ions diffuse and positive when sodium ions diffuse because of opposite concentration gradients of these two ions

or potassium. In Figure 10–1A the potassium concentration is very great inside the membrane while that outside is very low Furthermore, the membrane is very permeable to the potassium ions but not to any other ions. Because of the large potassium concentration gradient from the inside toward the outside, there is a strong tendency for potassium ions to diffuse outward. As they do so, they carry positive charges to the outside, thus creating a state of electropositivity outside the membrane, and electronegativity on the inside because of the negative anions that remain behind, which do not diffuse outward along with the potassium. This potential difference across the membrane tends to repel the potassium ions in a backward direction from outside toward the inside. And within a millisecond or so the potential becomes great enough to block further diffusion of potassium ions to the exterior. The potential at this point is the *Nernst potential for the potassium ions,* as we shall discuss.

Figure 10–1B illustrates the same effect as [illegible] in Figure 10–1A but with a high [illegible] sodium ions outside the membrane and [illegible]

sodium concentration inside. These ions are also positively charged. However, this time the membrane is highly permeable to the sodium ions and impermeable to all other ions. Diffusion of the sodium ions to the inside creates a membrane potential now of opposite polarity, with negativity outside and positivity inside, and, again, the membrane potential rises high enough within milliseconds to block further net diffusion of the sodium ions to the inside, but this time the potential is called the *Nernst potential for the sodium ions*

Thus, in both Figures 10–1A and B we see that a concentration difference of ions across a semipermeable membrane can, under appropriate conditions, cause the creation of a membrane potential. In later sections of this chapter, we shall also see that many of the membrane potential changes observed during the course of nerve and muscle impulse transmission result from the occurrence of rapidly changing diffusion membrane potentials of this nature

Relationship of the Diffusion Potential to the Concentration Difference—The Nernst Equation. When a concentration difference of a single type of ion across a membrane causes diffusion of ions through the membrane, thus creating a membrane potential, the magnitude of the potential *inside* the membrane versus the *outside* is determined by the ratio of the tendency for the ions to diffuse in one direction or the other direction, which is determined for positive ions by the following formula *(at body temperature, 38° C).*

$$\text{EMF (millivolts)} = -61 \log \frac{\text{Conc. inside}}{\text{Conc. outside}}$$

Thus, when the concentration of positive ions on the inside of a membrane is ten times that on the outside, the log of 10 is 1, and the potential difference calculates to be −61 millivolts. This equation is called the *Nernst equation,* and the developed potential is the *Nernst potential*

However, two conditions are necessary for this Nernst potential to develop as a result of diffusion: (1) The membrane must be selectively permeable, allowing a single type of ion to diffuse through the channels while all other ions do not diffuse (2) The concentration of the diffusible ions must be greater on one side of the membrane than on the other side.

Using the Nernst equation, let us now calculate the Nernst potential across the nerve membrane, first, when the membrane is permeable only to sodium ions and, second, when the membrane is permeable only to potassium ions:

The normal concentration of sodium ions inside the nerve membrane is approximately 14 mEq and outside approximately 142 mEq. Thus, the ratio of these two is 0.10, and the logarithm of 0.10 is −1.00. Multiplying this by −61 millivolts gives a Nernst potential for sodium of +61 millivolts inside the nerve fiber membrane.

The normal concentration of potassium ions inside the nerve fiber is approximately 140 mEq per liter and 4 mEq per liter on the outside. The ratio of these two is 35. The logarithm of 35 is 1.54; this times −61 millivolts equals a Nernst potential for potassium of −94 millivolts inside the membrane

Therefore, if there were no pumping of ions through the nerve membrane and if the membrane were permeable only to sodium but not at all to other ions, the potential inside the nerve fiber would be +61 millivolts Conversely, if the membrane were permeable to potassium but not permeable to any other ions, the membrane potential would be −94 millivolts. We shall see later in this discussion that under resting conditions the membrane potential averages about −90 millivolts which is very near the −94 millivolts potassium Nernst potential. This is true because in the resting state the membrane is very permeable to potassium and only slightly permeable to sodium On the other hand, when a nerve impulse is transmitted, the membrane, for a minute fraction of a second, becomes much more permeable to sodium than to potassium. Therefore, during this split second, the membrane potential rises to approximately +45 millivolts, which is much nearer the sodium Nernst potential than the potassium Nernst potential.

Calculation of the Diffusion Potential When the Membrane Is Permeable to Several Different Ions

When a membrane is permeable to several different ions, the diffusion potential that will develop depends on three factors: (1) the polarity of the electrical charge of each ion, (2) the permeability of the membrane *(P)* to each ion, and (3) the concentration *(C)* of the respective ions on the inside *(i)* and outside *(o)* of the membrane. Thus the following formula, called the *constant field equation,* the *Goldman equation,* or the *Goldman-Hodgkin-Katz equation,* gives the calculated membrane potential on the *inside* of the membrane when two univalent positive ions, sodium (Na^+) and potassium (K^+), and one univalent negative ion, chloride ion (Cl^-), are involved.

$$\text{EMF (millivolts)} = -61 \cdot \log \frac{C_{Na^+,i} P_{Na^+} + C_{K^+,i} P_{K^+} + C_{Cl^-,o} P_{Cl^-}}{C_{Na^+,o} P_{Na^+} + C_{K^+,o} P_{K^+} + C_{Cl^-,i} P_{Cl^-}}$$

Now, let us study the importance and the meaning of this equation. First, sodium, potassium, and chloride ions are the ions most importantly involved in the development of membrane potentials in nerve and muscle fibers, as well as in the neuronal cells in the central nervous system The

concentration gradient of each of these ions across the membrane helps determine the voltage of the membrane potential.

Second, the degree of importance of each of the ions in determining the voltage is proportional to the membrane permeability for that particular ion. Thus, if the membrane is impermeable to both potassium and chloride ions, then the membrane potential becomes entirely determined by the concentration gradient of sodium ions alone, and the resulting potential will be exactly equal to the Nernst potential for sodium. The same principle holds for each of the other two ions if the membrane should become selectively permeable for either one of them alone.

Third, a positive ion concentration gradient from *inside* the membrane *to the outside* causes electronegativity inside the membrane. The reason for this is that positive ions diffuse to the outside when their concentration is higher inside than outside. This carries positive charges to the outside but leaves the negative anions on the inside. Exactly the opposite effect occurs when there is a negative ion gradient. That is, a chloride ion gradient *from the outside to the inside* causes negativity inside the cell, because negatively charged chloride ions then diffuse to the inside while leaving the positive ions on the outside.

Fourth, we shall see later that the permeabilities of the sodium and potassium channels undergo very rapid changes during conduction of the nerve impulse, whereas the permeability of the chloride channels does not change greatly during this process. Therefore, the changes in the sodium and potassium permeabilities are primarily responsible for signal transmission in the nerves, which is the subject of most of the remainder of this chapter.

MEMBRANE POTENTIALS CAUSED BY ACTIVE TRANSPORT—THE SODIUM-POTASSIUM "ELECTROGENIC PUMP"

Figure 10–2 illustrates another method by which a membrane potential can develop—by active transport. It will be recalled from the discussions in the last chapter that the sodium-potassium pump pumps three sodium ions out of the cell for every two potassium ions pumped in. Thus, for each cycle of the pump the inside of the nerve fiber loses one positive charge. Because the membrane is not permeable to most of the negatively charged ions (anions) inside the cell, continuation of this process will lead to an excess of positive charges on the outside and an excess of negative charges inside, as illustrated in Figure 10–2. Thus, once again, the nerve membrane becomes negatively charged on the inside. And, because of the ability of the Na^+-K^+ pump to create such a membrane potential, it is called an *electrogenic pump*.

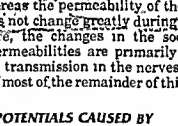


Figure 10–2. Establishment of a membrane potential as a result of sodium and potassium being pumped through the nerve membrane by the Na^+-K^+ electrogenic pump, with three sodium ions passing outward through the membrane for each two potassium ions passing inward

It was also clear from the discussions of the previous chapter that the Na^+-K^+ pump is responsible for establishing the normal sodium and potassium gradients across the cell membrane. That is, the continual pumping of sodium to the exterior leads to the normally greatly reduced sodium concentration inside the nerve fiber, whereas the pumping of the potassium ions to the interior helps establish the high concentration of potassium on the inside. We shall see later in the chapter that after many nerve impulses have been transmitted, the concentration gradients across the cell membrane of both sodium and potassium decrease because of diffusion of these ions through the membrane during the action potentials. But the Na^+-K^+ pump will soon re-establish the appropriate concentration gradients.

MEASURING THE MEMBRANE POTENTIAL

The method for measuring the membrane potential is simple in theory but often very difficult in practice because of the small sizes of many of the fibers. Figure 10–3 illustrates a small pipette filled with a very strong electrolyte solution (KCl) that is impaled through the cell membrane to the interior of the fiber. Then another electrode, called the "indifferent electrode," is placed in the interstitial fluids, and the potential difference between the inside and outside of the fiber is measured using an appropriate voltmeter. This is a highly sophisticated electronic apparatus that is capable of measuring *very small voltages despite extremely high resistance* to electrical flow through the tip of the micropipette.

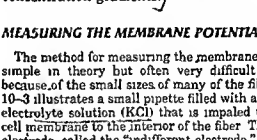


Figure 10–3. Measurement of the membrane potential of the nerve fiber using a microelectrode.

which has a diameter usually less than 1 micron and a resistance often as great as a billion ohms. For recording rapid *changes* in the membrane potential during the transmission of nerve impulses the microelectrode is connected to an oscilloscope, as will be explained later in the chapter.

THE CELL MEMBRANE AS AN ELECTRICAL CAPACITOR

In each of the figures shown thus far, the negative and positive ionic charges that cause the membrane potential have been shown to be lined up against the membrane, and we have not spoken of the arrangement of the charges elsewhere in the fluids, either inside the nerve fiber or on the outside in the interstitial fluid. However, Figure 10–4 illustrates this, showing that everywhere except adjacent to the surfaces of the cell membrane itself, the negative and positive charges are exactly equal. This is called the principle of *electrical neutrality*; that is, for every positive ion there is a negative ion nearby to neutralize it, or otherwise electrical potentials of billions of volts would appear within the fluids.

When positive charges are pumped to the outside of the membrane, these positive charges line up along the outside of the membrane, and on the inside the anions line up that have been left behind. This creates a *dipole layer* of positive and negative charges between the outside and inside of the membrane, but it still leaves equal numbers of negative and positive charges everywhere else within the fluids. This is the same effect that occurs when the plates of an electrical capacitor becomes electrically charged—that is, lining up of negative and positive charges on the opposite sides of the dielectric membrane between the plates. Therefore, the lipid bilayer of the cell membrane actually functions as a dielectric of a cell membrane capacitor, much as mica, paper, and Mylar function as dielectrics in electrical capacitors.

Because of the extreme thinness of the cell membrane (only 7 to 10 nanometers), its capacitance is tremendous for its area—about *1 microfarad per square centimeter.*

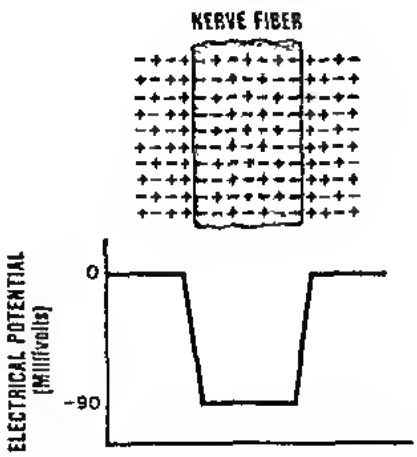


Figure 10–4. Distribution of positively and negatively charged ions in the interstitial fluid surrounding a nerve fiber and in the fluid inside the fiber; note the dipolar alignment of negative charges along the inside surface of the membrane and positive charges along the outside surface. In the lower panel are illustrated the abrupt changes in membrane potential that occur at the membranes on the two sides of the fiber.

The lower part of Figure 10–4 illustrates the electrical potential that will be recorded at each point in or near the nerve fiber membrane, beginning at the left side of the figure and passing to the right. As long as the electrode is outside the nerve membrane, the potential that is recorded is zero, which is the potential of the extracellular fluid. Then, as the recording electrode passes through the electrical dipole layer at the cell membrane, the potential decreases immediately to −90 millivolts. Again, the electrical potential remains at a steady level as the electrode passes across the interior of the fiber but reverses back to zero the instant it passes through the opposite side of the membrane.

The fact that the nerve membrane functions as a capacitor has one especially important point of significance. To create a negative potential inside the membrane, only enough positive ions must be transported outward to develop the electrical dipole layer at the membrane itself. All the remaining ions inside the nerve fiber can still be both positive and negative ions. Therefore, an incredibly small number of ions needs to be transferred through the membrane to establish the normal potential of −90 millivolts inside the nerve fiber—only about 1/500,000 of the total positive charges inside the fiber need be so transferred. Also, an equally small number of positive ions moving to the inside of the fiber can reverse the potential from −90 millivolts to as much as +35 millivolts within as little as 1/10,000 of a second. This rapid shifting of ions in this manner causes the nerve signals that we will discuss in the subsequent sections of this chapter.

THE RESTING MEMBRANE POTENTIAL OF NERVES

The membrane potential of large nerve fibers when they are not transmitting nerve signals—that is, when they are in the so-called "resting" state—is about −90 millivolts (mv). That is, the potential *inside the fiber* is 90 mv more negative than the potential in the interstitial fluid on the outside of the fiber. In the next few paragraphs we will explain how this potential comes about, but before doing so we must describe the transport properties of the resting nerve membrane for sodium and potassium.

Active Transport of Sodium and Potassium Ions Through the Membrane—The Sodium-Potassium Pump. First, let us recall from the discussions of the previous chapter that all cell membranes of the body have a powerful sodium-potassium pump and that this continually pumps sodium to the outside of the fiber and potassium to the inside. Further, let us remember that this is an *electrogenic pump*, because more positive charges are pumped to the outside than to the inside, leaving a net deficit of positive ions on the inside; this is the same as causing a negative charge inside the cell membrane.

This sodium-potassium pump causes the tremendous concentration gradients for sodium and po-

tassium across the resting nerve membrane. These gradients are the following:

Na^+ (outside):	142 mEq/L
Na^+ (inside):	14 mEq/L
K^+ (outside):	4 mEq/L
K^+ (inside):	140 mEq/L

And the ratios of these two respective ions from the inside to the outside are

$$Na^+_{inside}/Na^+_{outside} = 0.1$$
$$K^+_{inside}/K^+_{outside} = 35.0$$

Leakage of Potassium and Sodium Through the Nerve Membrane. To the right in Figure 10–5 is illustrated a channel type of transport protein in the cell membrane through which potassium and sodium ions can leak. The channels through these proteins are called *potassium-sodium "leak" channels.* The emphasis is on the potassium because these channels are far more permeable to potassium than to sodium, normally about 100 times times as permeable. We shall see later that this differential in permeability is exceedingly important in the establishment of the *normal resting membrane potential.*

ORIGIN OF THE NORMAL RESTING MEMBRANE POTENTIAL

Figure 10–6 illustrates the important factors in the establishment of the normal resting membrane potential of −90 mv. These are

Contribution of the Potassium Diffusion Potential. In Figure 10–6A, we make the assumption that the only movement of ions through the membrane is the diffusion of potassium ions, as illustrated by the open channel between the potassium inside the membrane and the outside. Because of the high ratio of potassium ions inside to the potassium ions outside, 35 to 1, the Nernst potential corresponding to this ratio is −94 mv. That is, the *diffusion potential* caused by the potassium

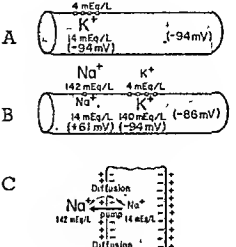


Figure 10–6. Establishment of resting membrane potentials in nerve fibers under three separate conditions. *A*, when the membrane potential is caused entirely by potassium diffusion alone; *B*, when the membrane potential is caused by diffusion of both sodium and potassium ions, and *C*, when the membrane potential is caused by diffusion of both sodium and potassium ions plus pumping of both these ions by the Na^+–K^+ pump.

would be −94 mv, as illustrated to the right inside the nerve fiber, if this were the only factor causing the resting nerve potential.

Contribution of Sodium Diffusion Through the Nerve Membrane. Figure 10–6B illustrates the addition of very slight permeability of the nerve membrane to sodium ions, caused by the minute diffusion of the sodium ions through the K^+-Na^+ leak channels. The ratio of sodium ions from inside to outside the membrane is 0.1, and this gives a calculated Nernst potential for the inside of the membrane of +61 mv. But also shown in Figure 10–6B is the Nernst potential for potassium diffusion of −94 mv. How do these interact with each other, and what will be the summated potential? This can be answered by using the Goldman equation described earlier. However, intuitively, one can see that if the membrane is highly permeable to potassium but only very slightly permeable to sodium, it is logical that the diffusion of potassium will contribute far more to the membrane potential than will the diffusion of sodium. In the normal nerve fiber, the permeability of the membrane to potassium is about 100 times as great as to sodium. Using this value in the Goldman equation gives an internal membrane potential of −86 mv, as shown to the right in the figure.

Contribution of the Na^+-K^+ Pump. Finally, in Figure 10–6C the additional contribution of the sodium-potassium pump is illustrated. In this fig-

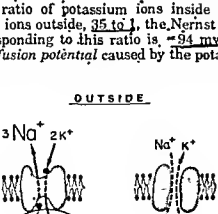


Figure 10–5. The functional characteristics of the Na^+–K^+ pump and also of the potassium-sodium "leak" channels.

ure, there is very slight diffusion of sodium ions, much greater diffusion of potassium ions, and continuous pumping of sodium ions to the outside and simultaneous pumping of potassium ions to the inside of the membrane, but with a ratio of sodium to potassium pumping of three to two, as discussed earlier. The fact that more sodium ions are being pumped to the outside than potassium to the inside causes a continual loss of positive charges from inside the membrane, and this creates an additional degree of negativity on the inside beyond that which can be accounted for by diffusion alone. Therefore, as illustrated in Figure 10-6C, the net membrane potential with all these factors operative at the same time is −90 mv.

In summary, the diffusion potentials alone caused by potassium and sodium diffusion would give a membrane potential of approximately −86 mv, almost all of this being determined by the potassium diffusion. Then, an additional −4 mv is contributed to the membrane potential by the electrogenic Na^+-K^+ pump, giving a net resting membrane potential of −90 mv.

The resting membrane potential in large skeletal muscle fibers is approximately the same as that in large nerve fibers, also −90 mv. However, in both small nerve fibers and small muscle fibers—smooth muscle, for instance—as well as in many of the neurons of the central nervous system, the membrane potential is often as little as −40 to −60 mv instead of −90 mv.

THE NERVE ACTION POTENTIAL

Nerve signals are transmitted by *action potentials*, which are rapid changes in the membrane potential. Each action potential begins with a sudden change from the normal resting negative potential to a positive membrane potential and then ends with an almost equally rapid change back again to the negative potential. To conduct a nerve signal, the action potential moves along the nerve fiber until it comes to the fiber's end. The upper panel of Figure 10-7 shows the disturbances that occur at the membrane during the action potential, with transfer of positive charges to the interior of the fiber at its onset and return of positive charges to the exterior at its end. The lower panel illustrates graphically the successive changes in the membrane potential over a period of a few 10,000ths of a second, illustrating the explosive onset of the action potential and the almost equally as rapid recovery.

The successive stages of the action potential are described by the following terms.

Resting Stage. This is the resting membrane potential before the action potential occurs. The membrane is said to be "polarized" during this stage because of the very large negative membrane potential that is present.

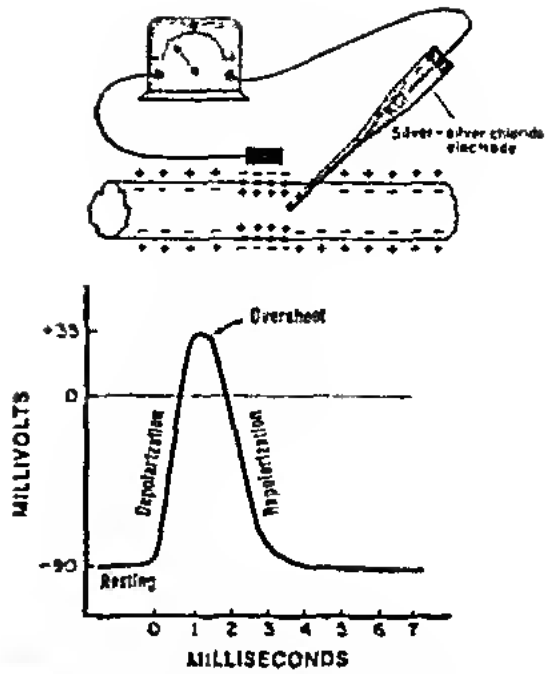


Figure 10-7. A typical action potential recorded by the method illustrated in the upper panel of the figure.

Depolarization Stage. At this time, the membrane suddenly becomes very permeable to sodium ions, allowing tremendous numbers of sodium ions to flow to the interior of the axon. The normal "polarized" state of −90 mv is lost, with the potential rising rapidly in the positive direction. This is called *depolarization*. In large nerve fibers, the membrane potential actually "overshoots" beyond the zero level and becomes slightly positive, but in some smaller fibers as well as many central nervous system neurons, the potential merely approaches the zero level and does not overshoot to the positive state.

Repolarization Stage. Within a few 10,000ths of a second after the membrane becomes highly permeable to sodium ions, the sodium channels close almost as rapidly as they had opened. Then, rapid diffusion of potassium ions to the exterior re-establishes the normal negative resting membrane potential. This is called *repolarization* of the membrane.

To explain more fully the factors that cause both the depolarization and repolarization processes, we need now to describe the special characteristics of yet two other types of transport channels through the nerve membrane: the voltage-gated sodium and potassium channels.

THE VOLTAGE-GATED SODIUM AND POTASSIUM CHANNELS

The principal actor in causing both depolarization and repolarization of the nerve membrane during the action potential is the *voltage-gated sodium channel*. However, the *voltage-gated potas-*

sium channel also plays an important role in some nerve fibers for the rapid repolarization of the membrane. *These two voltage-gated channels are in addition to the Na^+-K^+ pump and also in addition to the Na^+-K^+ leak channels.*

The Voltage-Gated Sodium Channel—"Activation" and "Inactivation" of the Channel

The upper panel of Figure 10–8 illustrates the voltage-gated sodium channel in three separate states. This channel has two separate gates, one near the outside of a channel called the *activation gate* and another near the inside called the *inactivation gate.* To the left is shown the state of these two gates in the normal resting membrane when the membrane potential is −90 mv. In this state the activation gate is closed, which prevents any entry of sodium ions to the interior of the fiber through these sodium channels. On the other hand, tha inactivation gate is open and does not at this time constitute any barrier to the movement of sodium ions.

Activation of the Sodium Channel. When the membrane potential becomes less negative than during the resting state, rising from −90 mv toward zero, it finally reaches a voltage, usually somewhere between −70 and −50 mv, that causes a sudden conformational change in the activation gate, flipping it to the open position. This is called the *activated state,* and during this state sodium ions can literally pour inward through the channel, increasing the sodium permeability of the membrane as much as 500- to 5000-fold.

Inactivation of the Sodium Channel. To the far right in the upper panel of Figure 10–8 is illustrated a final state of the sodium channel. The same increase in voltage that opens the activation gate also closes the inactivation gate. However, closure of the inactivation gate occurs a few 10,000ths of a second after the activation gate opens. That is, the conformational change that flips the inactivation gate to the closed state is a slower process, while the conformational change that opens the activation gate is a very rapid process. Therefore, after the sodium channel has remained open for a few 10,000ths of a second, it suddenly closes, and sodium ions can no longer pour to the inside of the membrane. At this point the membrane potential begins to recover back toward the resting membrane state, which is the repolarization process.

A very important characteristic of the sodium channel inactivation process is that *the inactivation gate will not reopen again until the membrane potential returns either to or nearly to the original resting membrane potential level.* Therefore, it is not possible for the sodium channels to open again without the nerve fiber first repolarizing.

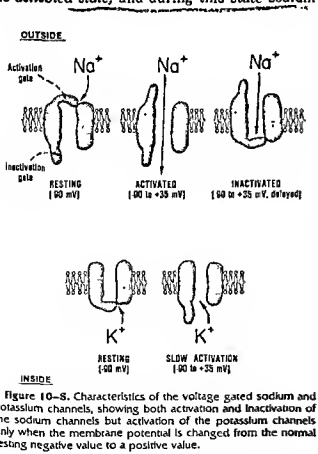


Figure 10–8. Characteristics of the voltage gated sodium and potassium channels, showing both activation and inactivation of the sodium channels but activation of the potassium channels only when the membrane potential is changed from the normal resting negative value to a positive value.

The Voltage-Gated Potassium Channels and Their Activation

The lower panel of Figure 10–8 illustrates the voltage-gated potassium channel in two separate states: during the resting membrane state and during the action potential. During the resting state, the gate of the potassium channel is closed, as illustrated to the left in the figure, and potassium ions are prevented from passing to the exterior. When the membrane potential rises from −90 mv toward zero, this voltage change causes a slow conformational opening of the gate and allows increased potassium diffusion outward through the membrane. However, because of the slowness of opening of these potassium channels, they open at the same time that the sodium channels are becoming inactivated and therefore are closing. Thus, the decrease in sodium entry to the cell and simultaneous increase in potassium exit from the cell greatly speeds the repolarization process, leading within a few 10,000ths of a second to full recovery of the resting membrane potential.

The Research Method for Measuring the Effect of Voltage on Opening and Closing of the Voltage-Gated Channels—The Voltage Clamp. The original research that led to our quantitative understanding of the sodium and potassium channels was so ingenious that it led to Nobel prizes for the scientists responsible, Hodgkin and Huxley. The essence of these studies is illustrated in Figures 10–9 and 10–10.

Figure 10–9 illustrates the experimental apparatus called the *voltage clamp,* which is used to measure the

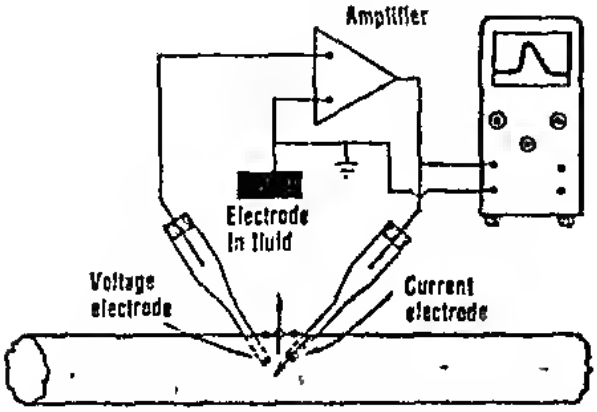


Figure 10–9. The "voltage clamp" method for studying flow of ions through specific channels.

flow of ions through the different channels Let us explain the voltage clamp Two electrodes are inserted into tha nerve fiber One of these is for the purpose of measuring the voltage of the membrane potential The other is to conduct electrical current either into or out of tha nerve fiber. This apparatus is used in the following way. The investigator decides what voltage he or she wishes to establish inside the nerve fiber He then adjusts the electronic portion of the apparatus to the desired voltage, and this automatically injects either positive or negative electricity through the current electrode at whatever rate is required to hold the voltage, as measured by the voltage electrode, at the level set by the operator For instance, when the membrane potential is suddenly increased by this voltage clamp from −90 mv to zero, the voltage-gated sodium and potassium channels open, and sodium and potassium ions begin to pour through the channels Furthermore, electrical current is injected through the current electrode of the voltage clamp to maintain the intracellular voltage at tha zero level To achieve this, the current injerted through the current electrode must be exactly equal to but of opposite polarity to the net current flow through the membrane channels To measure how much current flow is occurring at each instant, the current electrode is connected to an oscilloscope that records the flow, as illustrated on the screen of the oscilloscope in the figure

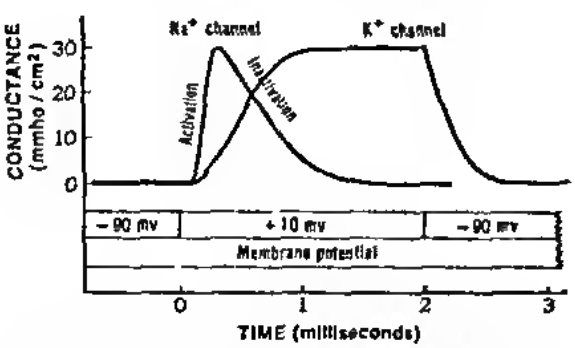


Figure 10–10. Changes in conductance of sodium and potassium ion channels when the membrane potential is suddenly increased from the normal resting value of −90 mv to a positive value of +10 mv for 2 milliseconds. This figure illustrates that the sodium channels open (activate) and then close (inactivate) before the end of the 2 milliseconds whereas the potassium channels only open (activate)

Finally, the investigator adjusts the concentrations of the ions to desired levels both inside and outside the nerve fiber and repeats the study. This can be done very easily when using very large nerve fibers removed from some crustaceans, especially the giant squid axon that is sometimes as large as 1 mm in diameter. When sodium is the only permeant ion in the solutions inside and outside the squid axon, then the voltage clamp will measure current flow only through the sodium channels When potassium is the only permeant ion, then current flow only through the potassium channels will be measured

Another means for studying the flow of ions through individual channels is to block one type of channel at a time. For instance, the sodium channels can be blocked by a toxin called *tetrodotoxin* when this is applied to the outside of the rell membrene where the sodium activation gates are located Conversely, *tetraethylammonium ion* will block the potassium pores when it is applied to the interior of the nerve fiber

Figure 10–10 illustrates the changes in conductance of the voltage-gated sodium and potassium channels when the membrane potential is suddenly changed by use of the voltage clamp from −90 mv to +10 mv and 2 milliseconds later back again to −90 mv. Note the sudden opening of the sodium channels (the activation stage) within a very small fraction of a millisecond after the membrene potential was increased to the positive value However, during approximately the next millisecond, the sodium channels automatically closed (the inactivation stage) and were fully closed within about 1 5 milliseconds after the onset of the positive membrane potential Thus, long before the membrane potential returned to its original resting value, the sodium channels had already completely closed, or "inactivated"

Now, note the difference in the opening (activation) of the potassium channels These opened slowly, reaching the full open state only after the sodium channels had already become almost completely closed Furthermore, once the potassium channels opened, they remained open for the entire duration of the positive membrane potential and did not close again until after the membrane potential was decreased back to −90 mv.

Finally, let us recall that the voltage-gated channels normally flip to the open state or the closed state very suddenly, which was illustrated in Figure 9–5 in the previous chapter Therefore, how is it that the curves in Figure 10–10 are so smooth? The answer is that these curves represent the flow of sodium and potassium ions through literally thousands of channels at the same time Some open at one voltage level, others at another voltage, and so forth Likewise, some become inactivated at different points in the cycle from the others Thus, the illustrated curves represent summated current flows through the many channels

SUMMARY OF THE EVENTS THAT CAUSE THE ACTION POTENTIAL

Figure 10–11 illustrates in summary form the sequential events that occur during and shortly after the action potential These are:

At the bottom of the figure are shown the changes in membrane conductances for sodium and potassium ions. During the resting state, before the action potential begins, the conductanre for

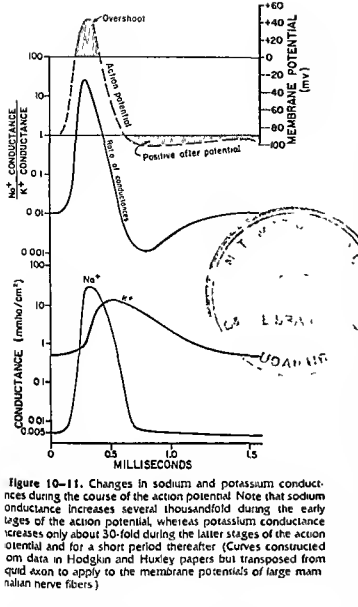


Figure 10–11. Changes in sodium and potassium conductances during the course of the action potential Note that sodium conductance increases several thousandfold during the early stages of the action potential, whereas potassium conductance increases only about 30-fold during the latter stages of the action potential and for a short period thereafter (Curves constructed from data in Hodgkin and Huxley papers but transposed from squid axon to apply to the membrane potentials of large mammalian nerve fibers)

potassium ions is shown to be already 50 to 100 times as great as the conductance for sodium ions. This is caused by much greater leakage of potassium ions through the leak channels than the leakage of sodium ions However, at the onset of the action potential, the sodium channels instantaneously become activated and allow an up to 5000-fold increase in sodium conductance. Then the inactivation process almost as rapidly closes the sodium channels within another few fractions of a millisecond. The onset of the action potential also causes voltage gating of the potassium channels, causing them to begin opening a fraction of a millisecond after the sodium channels open. And at the end of the action potential, the return of the membrane potential to the negative state causes the potassium channels to close back to their original status, but again only after a short delay.

In the middle portion of Figure 10–11 is shown the ratio of sodium conductance to potassium conductance at each instant during the action potential, and above this is shown the action potential itself. During the early portion of the action potential, this ratio increases more than a thousandfold. Therefore, far more sodium ions now flow to the interior of the fiber than do potassium ions to the exterior This is what causes the membrane potential to become positive. At this time the sodium channels become inactivated but at the same time the potassium channels open, and the ratio of conductance now shifts far in favor of high potassium conductance but low sodium conductance. This allows extremely rapid loss of potassium ions to the exterior while essentialy no sodium ions flow to the interior. Consequently, the action potential quickly returns to its baseline level

The Positive After-Potential

Note also in Figure 10–11 that the membrane potential becomes even more negative than the original resting membrane potential for a few milliseconds after the action potential is over This is called the *positive after-potential*, which is really a misnomer because the positive after-potential is actually somewhat more negative than the *resting potential* However, the reason for calling it "positive" is that, historically, the first potential measurements were made on the outside of the nerve fiber membrane rather than inside, and when measured on the outside this potential caused a positive record on the meter rather than a negative one.

The cause of the positive after potential is mainly that many potassium channels remain open for several milliseconds after the repolarization process of the membrane is complete This allows excess potassium ions to diffuse out of the nerve fiber, leaving an extra deficit of positive ions on the inside, which means more negativity

ROLES OF OTHER IONS DURING THE ACTION POTENTIAL

Thus far, we have considered only the roles of sodium and potassium ions in the generation of the action potential However, at least three other types of ions must be considered These are

The Impermeant Negatively Charged Ions (Anions) Inside the Axon. Inside the axon are many negatively charged ions that cannot go through the membrane channels. These include protein molecules, many organic phosphate compounds, sulfate compounds, and so forth. Since these cannot leave the interior of the axon, any deficit of positive ions inside the membrane leaves an excess of the impermeant negative ions. Therefore, these impermeant negative ions are responsible for the negative charge inside the fiber.

Calcium Ions. The cell membranes of almost all, if not all, cells of the body have a calcium pump similar to the sodium pump. And, like the sodium pump, this pump pumps calcium ions from the interior to the exterior of the cell membrane, creating a calcium ion gradient of about 10,000-fold, leaving an internal concentration of calcium ions of about 10^{-7} molar in contrast to an external concentration of about 10^{-3} molar.

In addition, there are also voltage-gated calcium channels. These channels are slightly permeable to sodium ions as well as to calcium ions; when they open both calcium and sodium ions flow to the interior

Therefore, these channels are frequently called *Ca^{++}-Na^{+} channels* The Ca^{++}-Na^{+} channels are very slow to become activated, requiring 10 to 20 times as long for activation as the normal sodium channels. Therefore, they are also frequently called *slow channels*, in contrast to the normal sodium channels that are called *fast channels*

The Ca^{++}-Na^{+} channels are very numerous in both cardiac muscle and smooth muscle. In fact, in some types of smooth muscle, the fast sodium channels are hardly present at all, so the action potentials then are caused almost entirely by activation of the slow Ca^{++}-Na^{+} channels

Increased Permeability of the Sodium Channels When There Is a Deficit of Calcium Ions. The concentration of calcium ions in the interstitial fluid also has a profound effect on the voltage level at which the sodium channels become activated. When there is a deficit of calcium ions, the sodium channels can be activated by very little increase of the membrane potential above the normal resting level. Therefore, the nerve fiber becomes highly excitable, sometimes discharging repetitively without any provocation rather than remaining in the resting state. In fact, the calcium ion concentration needs to fall only 30 to 50 per cent below normal before spontaneous discharge occurs in many peripheral nerves, often causing muscle "tetany" that can actually be lethal because of tetanic contraction of the respiratory muscles On the other hand, an increase in calcium ion concentration above normal decreases the excitability of the nerve fiber for opposite reasons.

The probable way in which calcium ions affect the sodium chanels is the following: These ions appear to bind to the exterior surfaces of the sodium channel protein molecule The positive charges of these calcium ions, in turn, alter the electrical state of the channel protein itself, in this way increasing the voltage level required to open the gate

Chloride Ions. Chloride ions leak through the resting membrane in the same way that small quantities of potassium and sodium ions leak through In the usual nerve fiber, the rate of chloride diffusion through the membrane is about one half as great as the diffusion of potassium ions Therefore, the question must be asked why have we not considered the chloride ions in our explanation of the action potential? The answer is that the chloride ions function passively in this process The permeability of the chloride leak channels does not change significantly during the action potential

In the normal resting state of the nerve fiber, the −90 mv inside the fiber repels most of the chloride ions from the fiber Therefore, the concentration of chloride ions inside the fiber is only 3 to 4 mEq per liter, whereas the concentration outside the fiber is about 103 mEq per liter The Nernst potential for this ratio of chloride ions is exactly equal to the −90 mv membrane potential, which is what one would expect for an ion that is not actively pumped

During the action potential, small quantities of chloride ions do diffuse into the nerve fiber because of the temporary loss of the internal negativity. This movement of the chloride ions serves to alter slightly the timing of the successive voltage changes during the action potential, but it does not alter the fundamental process

INITIATION OF THE ACTION POTENTIAL

Up to this point, we have explained the changi sodium and potassium permeabilities of the mer brane as well as the development of the actic potential itself, but we have not explained wh initiates the action potential. The answer to th is really quite simple:

First, as long as the membrane of the nerve fib remains totally undisturbed, no action potenti will occur in the normal nerve. However, if a event at all causes enough initial rise in tl membrane potential from −90 mv up toward tl zero level, the rising voltage itself will direct affect the voltage-gated sodium channels, causi them to become activated (opened). This allo rapid inflow of sodium ions, which causes st further rise of the membrane potential, thus ope ing still more voltage-gated sodium channels a more streaming of sodium ions to the interior the fiber, with the process continuing in a positiv feedback vicious circle until all the voltage-gat sodium channels have become totally activate But then within another fraction of a millisecor the rising membrane potential causes inactivatk of the sodium channels, and the action potenti soon terminates.

Threshold for Initiation of the Action Pote tial. An action potential will not occur until tl initial rise in membrane potential is great enou to create the vicious circle described in the la paragraph. Usually, a sudden rise in membra potential of 15 to 30 mv is required. Therefore, sudden increase in the membrane potential in large nerve fiber of from −90 mv up to about −f mv will usually cause the explosive developme of the action potential. This level of −65 mv said to be the *threshold* for stimulation

Accommodation of the Membrane—Failure Fire Despite Rising Voltage. If the membrane pote tial rises very slowly—over many milliseconds inste of a fraction of a millisecond—the slow, inactivatir gates of the sodium channels will then have time close at the same time that the activating gates a opening Consequently, the opening of the activatir gates will not be as effective in increasing the flow sodium ions as normally because many of the inactiva ing gates will already be closed Therefore, a slo increase in the internal potential of a nerve fiber eithe will require a higher threshold voltage than normal cause firing or will prevent firing entirely at times, eve with a voltage rise all the way to zero or even to positiv voltage This phenomenon is called *accommodation* the membrane to the stimulus.

PROPAGATION OF THE ACTION POTENTIAL

In the preceding paragraphs we have discusse the action potential as it occurs at one spot on th

membrane. However, an action potential elicited at any one point on an excitable membrane usually excites adjacent portions of the membrane, resulting in propagation of the action potential. The mechanism of this is illustrated in Figure 10–12. Figure 10–12A shows a normal resting nerve fiber, and Figure 10–12B shows a nerve fiber that has been excited in its midportion—that is, the midportion has suddenly developed increased permeability to sodium. The arrows illustrate a "local circuit" of current flow between the depolarized areas of the membrane and the adjacent resting membrane areas; positive electrical charges flow inward through the depolarized membrane and then for several millimeters along the core of the axon. These positive charges increase the voltage for a distance of 1 to 3 mm inside large fibers to above the threshold voltage value for initiating an action potential. Therefore, the sodium channels in these new areas immediately activate, and, as illustrated in Figure 10–12C and D, the explosive action potential spreads. And these newly depolarized areas cause local circuits of current flow still further along the membrane, causing progressively more and more depolarization. Thus, the depolarization process travels in both directions along the entire extent of the fiber. The transmission of the depolarization process along a nerve or muscle fiber is called *a nerve or muscle impulse.*

Direction of Propagation. It is now obvious that an excitable membrane has no single direction of propagation, but that the action potential can travel in both directions away from the stimulus—and even along all branches of a nerve fiber—until the entire membrane has become depolarized.

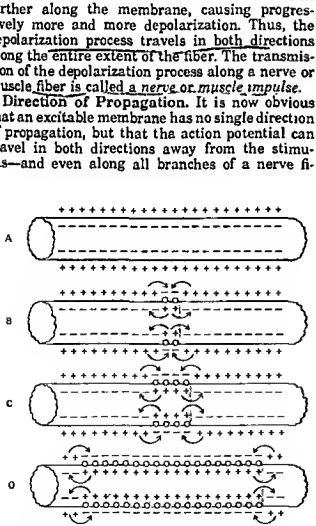


Figure 10–12. Propagation of action potentials in both directions along a conductive fiber.

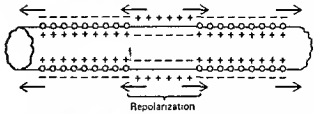


Figure 10–13. Propagation of repolarization in both directions along a conductive fiber

The All-or-Nothing Principle. It is equally obvious that, once an action potential has been elicited at any point on the membrane of a normal fiber, the depolarization process will travel over the entire membrane. This is called the all-or-nothing principle, and it applies to all normal excitable tissues. Occasionally, though, when the fiber is in an abnormal state the action potential will reach a point on the membrane at which it does not generate sufficient voltage to stimulate the adjacent area of the membrane. When this occurs the spread of depolarization will stop. Therefore, for normal propagation of an impulse to occur, the ratio of action potential to threshold for excitation, called the *safety factor,* must at all times be greater than 1

Propagation of Repolarization. The action potential normally lasts almost the same length of time at each point along a fiber. Therefore, repolarization normally occurs first at the point of original stimulus and then spreads progressively along the membrane, moving in the same direction that depolarization had previously spread. Figure 10–13 illustrates the same nerve fiber as that in Figure 10–12, showing that the repolarization process is propagated in the same direction as the depolarization process but a few 10,000ths of a second later.

"RECHARGING" THE FIBER MEMBRANE AFTER ACTION POTENTIALS—IMPORTANCE OF ENERGY METABOLISM

Transmission of each impulse along the nerve fiber reduces the concentration differences of sodium and potassium between the inside and outside of the membrane because of diffusion of sodium ions to the inside during depolarization and diffusion of potassium ions to the outside during repolarization. For a single action potential, this effect is so minute that it cannot even be measured. Indeed, 100,000 to 500,000 impulses can be transmitted by nerve fibers before the concentration differences have run down to the point that action

potential conduction ceases. Yet, even so, with time it becomes necessary to re-establish the sodium and potassium membrane concentration differences. This is achieved by the action of the Na^+-K^+ pump in exactly the same way as that described earlier in the chapter for establishment of the original resting potential. That is, the sodium ions that, during the action potentials, have diffused to the interior of the cell and the potassium ions that have diffused to the exterior are returned to their original state by the sodium and potassium pump. Since this pump requires energy for operation, this process of "recharging" the nerve fiber is an active metabolic one, utilizing energy derived from the adenosine triphosphate energy "currency" system of the cell

A special feature of the sodium-potassium ATPase membrane pumping system is that its degree of activity is very strongly stimulated by excess sodium ions inside the cell membrane. In fact, the pumping activity increases approximately in proportion to the third power of the sodium concentration That is, if the internal sodium concentration rises from 10 to 20 mEq per liter, the activity of the pump does not merely double but instead increases approximately eightfold. Therefore, it can easily be understood how the recharging process of the nerve fiber can rapidly be set into motion whenever the concentration differences of sodium and potassium across the membrane begin to "run down"

Heat Production by the Nerve Fiber. Figure 10-14 illustrates the relationship of heat production in a nerve fiber to the number of impulses transmitted by the fiber each second The rate of heat production is a measure of the rate of metabolism in the nerve, because heat is always liberated as a product of the chemical reactions of energy metabolism Note that the heat production increases markedly as the number of impulses per second increases It is this increased use of energy that causes the "recharging" process

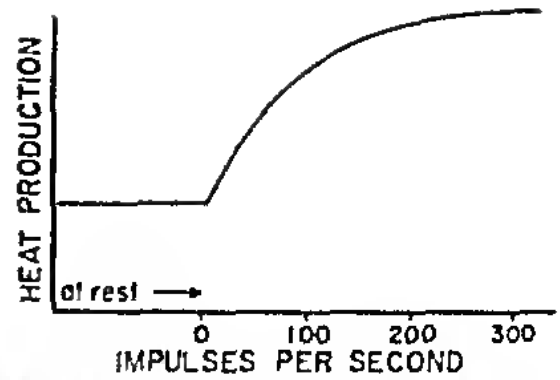


Figure 10-14. Heat production in a nerve fiber at rest and at progressively increasing rates of stimulation

THE SPIKE POTENTIAL AND THE AFTER-POTENTIALS

Figure 10-15 illustrates an action potential recorded with a much slower time scale than that illustrated in Figure 10-7, many milliseconds of recording are shown in comparison with only the first millisecond of the action potential in Figure 10-7

The Spike Potential. The initial, very large change in membrane potential shown in Figure 10-15 is called the spike potential In large, type A myelinated nerve fibers it lasts for about 0 4 millisecond The spike potential is the same as the action potential that has been discussed in the preceding paragraphs and is also called the *nerve impulse*

The Negative After-Potential. At the termination of the spike potential, the membrane potential sometimes fails to return all the way to its resting level for another few milliseconds, as shown in Figure 10-15. This is particularly likely to occur after a series of rapidly repeated action potentials. This is called the *negative after-potential*. It is believed to result at least partly from a buildup of potassium ions immediately outside the membrane, this causes the concentration ratio of potassium across the membrane to be temporarily less than normal and therefore prevents full return of the normal resting membrane potential for a few additional milliseconds

The Positive After-Potential. Once the membrane potential has returned to its resting value, it then becomes a little more negative than its normal resting value, this is the *positive after-potential* that was discussed earlier It is a fraction of a millivolt to a few millivolts more negative than the normal resting membrane potential, but it can last from 50 milliseconds to as long as many seconds

The first part of this positive after potential is caused by the excess permeability of the nerve membrane to potassium ions at the end of the spike potential, as discussed earlier However, the prolonged continuance of this potential is caused principally by the electrogenic pumping of excess sodium outward through the nerve fiber membrane, which is the recharging process that was discussed previously If the active transport processes are poisoned, this part of the positive after-potential is lost.

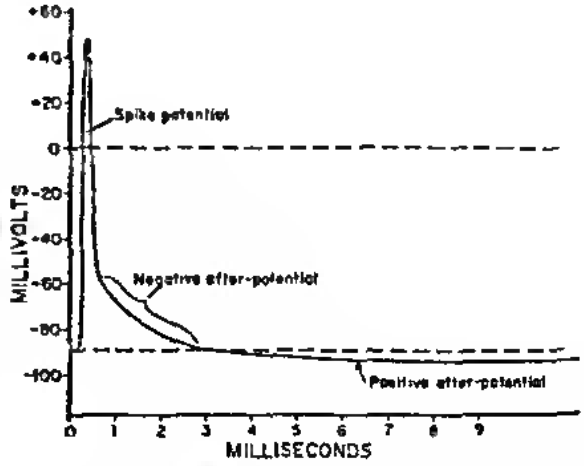


Figure 10-15. An idealized action potential, showing the initial spike followed by a negative after potential and a positive after-potential

PLATEAU IN SOME ACTION POTENTIALS

In some instances the excitable membrane does not repolarize immediately after depolarization, but, instead, the potential remains on a plateau near the peak of the spike sometimes for many milliseconds before repolarization begins. Such a plateau is illustrated in Figure 10–16, from which one can readily see that the plateau greatly prolongs the period of depolarization. This type of action potential occurs in the heart, where the plateau lasts for as long as three- to four-tenths second and causes contraction of the heart muscle during this entire period of time.

The cause of the action potential plateau is a combination of several different factors. First, in heart muscle, two separate types of channels enter into the depolarization process: (1) the usual voltage-activated sodium channels, called the *fast channels*, and (2) the voltage-activated calcium-sodium channels, which are slow to be activated and therefore are called *slow channels*—these chennels allow diffusion mainly of calcium ions but also of some sodium ions as well. Activation of the fast channels causes the spike portion of the action potential, whereas the slow but prolonged activation of the slow channels is mainly responsible for the plateau portion of this type of action potential.

A second factor partly responsible for the plateau is that the voltage-gated potassium channels are slow to be activated in some excitable tissues, often not opening until the very end of the plateau. This delays the return of the membrane potential toward the resting value. But, then, this opening of the potassium channels at the same time that the slow channels begin to close at the end of the plateau causes rapid return of the action potential from its plateau level back to the negative resting level.

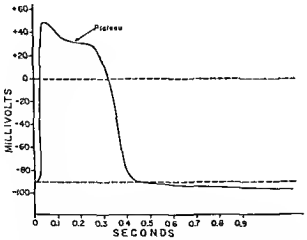


Figure 10–16. An action potential from a Purkinje fiber of the heart, showing a "plateau."

RHYTHMICITY OF CERTAIN EXCITABLE TISSUES—REPETITIVE DISCHARGE

All excitable tissues can discharge repetitively if the threshold for stimulation is reduced low enough. For instance, even nerve fibers and skeletal muscle fibers, which normally are highly stable, discharge repetitively when they are placed in a solution containing the drug veratrine or when the calcium ion concentration falls below a critical value. Repetitive discharges, or rhythmicity, occur normally in the heart, in most smooth muscle, and also in many of the neurons of the central nervous system. It is these rhythmical discharges that cause the heart beat, that cause peristalsis, and that cause such neuronal events as the rhythmical control of breathing.

The Re-Excitation Process Necessary for Rhythmicity. For rhythmicity to occur, the membrane, even in its natural state, must already be permeable enough to sodium ions (or to calcium and sodium ions through the Ca^{++}-Na^{+} slow channels) to allow automatic membrane depolarization. Thus, Figure 10–17 shows that the "resting" membrane potential is only −60 to −70 millivolts. This is not enough negative voltage to keep the sodium and calcium-sodium channels closed. That is, (a) sodium and calcium ions flow inward, (b) this further increases the membrane permeability, (c) still more ions flow inward, (d) the permeability increases more, and so forth, thus eliciting the regenerative process of sodium and calcium-sodium channel openings until an action potential is generated. Then, at the end of the action potential the membrane repolarizes. But shortly thereafter, the depolarization process begins again and a new action potential occurs spontaneously—this cycle continuing again and again and causing self-induced rhythmical excitation of the excitable tissue.

Yet, why does the membrane not depolarize immediately after it has become repolarized rather than delaying for nearly a second before the onset of the next action potential? The answer to this can be found by referring back to Figure 10–11, which shows that toward the end of all action

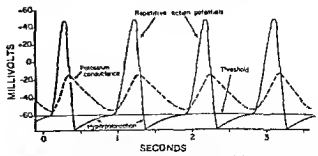


Figure 10–17. Rhythmic action potentials, and their relationship to potassium conductance and to the state of hyperpolarization.

potentials, and continuing for a short period thereafter, the membrane becomes excessively permeable to potassium. The excessive outflow of potassium ions carries tremendous numbers of positive charges to the outside of the membrane, creating inside the fiber considerably more negativity than would otherwise occur for a short period after the preceding action potential is over, thus drawing the membrane potential nearer to the potassium Nernst potential. This is a state called *hyperpolarization*, which is illustrated in Figure 10–17. As long as this state exists, re-excitation will not occur; but gradually the excess potassium conductance (and the state of hyperpolarization) disappears, thereby allowing the membrane potential to increase until it reaches the *threshold* for excitation; then suddenly a new action potential results, the process occurring again and again.

SPECIAL ASPECTS OF SIGNAL TRANSMISSION IN NERVE TRUNKS

Myelinated and Unmyelinated Nerve Fibers. Figure 10–18 illustrates a cross-section of a typical small nerve trunk, showing a few very large nerve fibers that compose most of the cross-sectional area and many more small fibers lying between the large ones. The large fibers are *myelinated* and the small ones are *unmyelinated*. The average nerve trunk contains about twice as many unmyelinated fibers as myelinated fibers.

Figure 10–19 illustrates a typical myelinated fiber. The central core of the fiber is the *axon*, and the membrane of the axon is the actual *conductive membrane*. The axon is filled in its center with *axoplasm*, which is a viscid intracellular fluid. Surrounding axon is a *myelin sheath* that is often thicker than axon itself, and about once every millimeter along extent of the axon the myelin sheath is ' .. . a *node of Ranvier*.

The myelin sheath is deposited around the axon b Schwann cells in the following manner: The membrar of a Schwann cell first envelops the axon. Then the ce rotates around the axon many times, laying down mu tiple layers of cellular membrane containing the lip substance *sphingomyelin*. This substance is an excelle insulator that prevents almost all flow of ions. In fac it increases the resistance to ion flow through the mer brane approximately 5000-fold and also decreases th membrane capacitance as much as fiftyfold. Howeve at the juncture between each two successive Schwar cells along the axon, a small, uninsulated area reman where ions can still flow with ease between the extr cellular fluid and the axon. This area is the node Ranvier.

"Saltatory" Conduction in Myelinated Fibe from Node to Node. Even though ions cannot flo significantly through the thick myelin sheaths of my linated nerves, they can flow with considerable ea through the nodes of Ranvier. Therefore, action pote tials can occur only at the nodes. Yet, the action pote tials are conducted from node to node, as illustrated Figure 10–20, this is called *saltatory conduction*. Th is, electrical current flows through the surroundi extracellular fluids and also through the axoplasm fro node to node, exciting successive nodes one after a other. Thus, the nerve impulse jumps down the fibe which is the origin of the term "saltatory."

Saltatory conduction is of value for two reasons: Firs by causing the depolarization process to jump lon intervals along the axis of the nerve fiber, this mech nism increases the velocity of nerve transmission

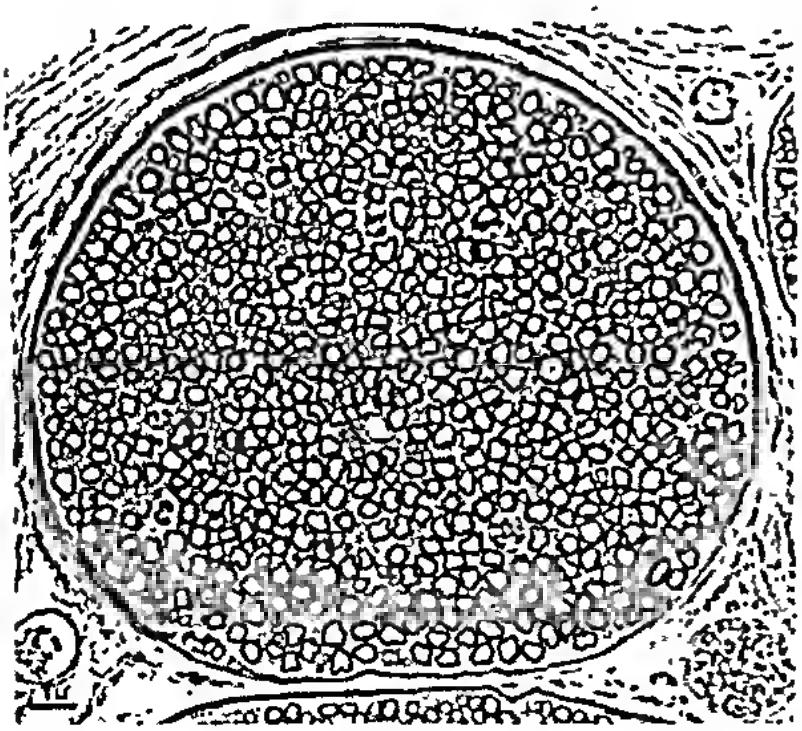
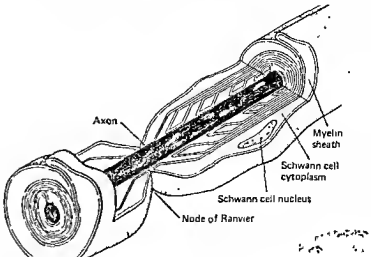
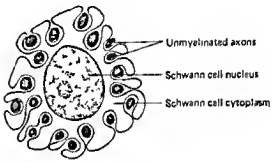


Figure 10–18. Cross section of a small nerv trunk containing myelinated and unmyelinated fi bers.



A



B

Figure 10–19. Function of the Schwann cell to insulate nerve fibers. A. The wrapping of a Schwann cell membrane around a large axon to form the myelin sheath of the myelinated nerve fiber (Modified from Leeson and Leeson, Atlas of Histology [Philadelphia, W. B. Saunders Company, 1979]. B. Evagination of the membrane and cytoplasm of a Schwann cell around multiple unmyelinated nerve fibers.

myelinated fibers an average of five- to seventfold. Second, saltatory conduction conserves energy for the axon, for only the nodes depolarize, allowing perhaps a hundred times smaller loss of ions than would otherwise be necessary and therefore requiring little extra metabolism for re-establishing the sodium and potassium concentration differences across the membrane after a series of nerve impulses.

Still another feature of saltatory conduction in large myelinated fibers is the following: The excellent insulation afforded by the myelin membrane and the fifty-fold decrease in membrane capacitance allows the repolarization process to occur with very little transfer of ions.

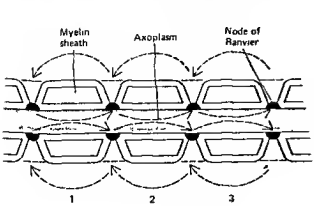


Figure 10–20. Saltatory conduction along a myelinated axon.

Therefore, at the end of the action potential when the sodium channels close, repolarization occurs so rapidly that the potassium channels usually have not yet begun to open significantly. Therefore, conduction of the nerve impulse in the myelinated nerve fiber is accomplished almost entirely by the sequential changes in the voltage-gated sodium channels, with very little contribution by the potassium channels

VELOCITY OF CONDUCTION IN NERVE FIBERS

The velocity of conduction in nerve fibers varies from as little as 0.5 meter per second in very small unmyelinated fibers to as high as 100 meters per second (the length of a football field) in very large myelinated fibers. The velocity increases approximately with the fiber diameter in myelinated nerve fibers and approximately with the square root of fiber diameter in unmyelinated fibers.

EXCITATION—THE PROCESS OF ELICITING THE ACTION POTENTIAL

Chemical Stimulation. Basically, any factor that causes sodium ions to begin to diffuse inward through the membrane in sufficient numbers will set off the automatic, regenerative opening of the sodium channels,

as noted earlier in the chapter, that eventuates in the action potential. Thus, certain chemicals can stimulate a nerve fiber by increasing the membrane permeability. Such chemicals include acids, bases, almost any salt solution of very strong concentration, and, most importantly, the substance *acetylcholine* Many nerve fibers, when stimulated, secrete acetylcholine at their endings where they synapse with other neurons or where they end on muscle fibers The acetylcholine in turn stimulates the successive neuron or muscle fiber by opening pores in the membrane with diameters of 0 6 to 0 7 nanometer, large enough for sodium (as well as other ions) to go through with ease This is discussed in much greater detail in Chapter 12, and it is one of the most important means by which nerve and muscle fibers are stimulated Likewise, *norepinephrine* secreted by sympathetic nerve endings can stimulate cardiac muscle fibers and some smooth muscle fibers, and still other hormonal *transmitter substances* can stimulate successive neurons in the central nervous system

Mechanical Stimulation. Crushing, pinching, or pricking a nerve fiber can cause a sudden surge of sodium influx and, for obvious reasons, can elicit an action potential Even slight pressure on some specialized nerve endings can stimulate these; this will be discussed in Chapter 48 in relation to sensory perception

Electrical Stimulation. An electrical current artificially induced through a nerve causes excess flow of ions through the axonal membrane; this too can initiate an action potential However, not all methods of applying electrical stimuli result in excitation and, since this is the usual means by which nerve fibers are excited when they are studied in the laboratory, the process of electrical excitation deserves more comment

Cathodal Versus Anodal Currents. If two electrodes are placed on a nerve bundle and direct electrical current is made to flow through the electrodes with one electrode positive and the other negative, the nerve fibers are stimulated at the negative electrode, called the "*cathode*" On the other hand, at the positive electrode, called the "*anode*," the nerve fibers actually become more resistant to excitation than normal Therefore, it is said that a *cathodal current* excites a nerve fiber while an *anodal current* inhibits it

The cause of these effects is the following Remember that the action potential is initiated by the opening of voltage-gated sodium channels Furthermore, these channels are opened by a decrease in the electrical voltage across the membrane The negative current from the cathode reduces the voltage immediately outside the membrane, drawing this voltage nearer to the voltage of the negative resting membrane potential inside the fiber This decreases the electrical voltage across the membrane and allows activation of the sodium channels, thus resulting in an action potential Conversely, at the anode, the injection of positive charges on the outside of the nerve membrane heightens the voltage difference across the membrane rather than lessening it. And this causes a state of "hyperpolarization," which decreases the excitability of the fiber.

Threshold for Excitation and "Acute Subthreshold Potential." A very weak electrical, mechanical, or chemical stimulus may not be able to excite a fiber. But, when this stimulus is progressively increased, there comes a point at which excitation takes place. Figure 10–21 illustrates the effects of successively applied stimuli of progressing strength A very weak stimulus at point A causes the membrane potential to change from −90 to −85 millivolts, but this is not sufficient change for the automatic regenerative processes of the action potential to develop At point B the stimulus is greater, but, here again, the intensity still is not enough to set off the automatic action potential. Nevertheless, the membrane voltage is disturbed for as long as a millisecond or more after both of the weak stimuli, the potential changes during these short intervals of time are called *acute subthreshold potentials*, as illustrated in the figure.

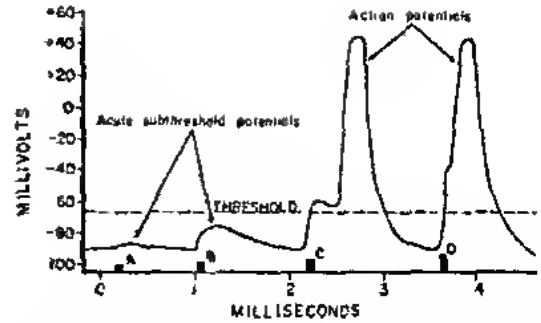


Figure 10–21. Effect of stimuli on the potential of the excitable membrane, showing the development of "acute sub-threshold potentials when the stimuli are below the threshold value required for eliciting an action potential

At point C in Figure 10–21 the stimulus elicits an acute membrane potential that is not subthreshold but slightly more than the threshold value, and, after a short "latent period," it initiates an action potential. At point D the stimulus is still stronger, and the acute membrane potential initiates the action potential even sooner. Thus, this figure shows that even a very weak stimulus always causes a local potential change at the membrane, but that the intensity of the *local potential* must rise to a *threshold value* before the automatic action potential will be set off

Excitability Curve of Nerve Fibers. A so-called "excitability curve" of a nerve fiber is shown in Figure 10–22 To obtain this curve a high voltage electrical stimulus (4 volts, in this instance) is applied to the fiber, and the minimum duration of stimulus required to excite the fiber is found The voltage and duration are plotted as point A. Then a stimulus voltage of 3 volts is applied,

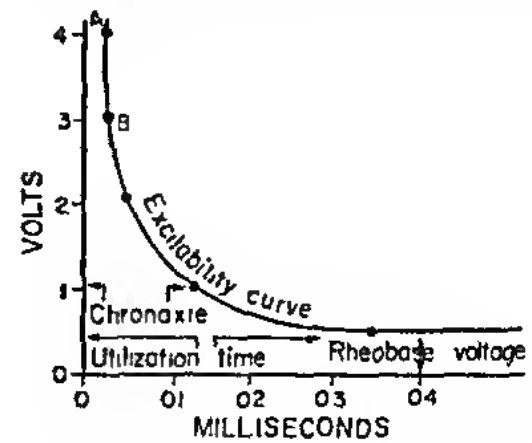


Figure 10–22. Excitability of a large myelinated nerve fiber

Greene, ... duration required is again determined; the ... are plotted as point B. The same is repeated at ... 1 volt, 0.5 volt, and so forth, until the least ... possible at which the membrane is stimulated ... seen reached On connection of these points, the ... excitability curve is generated.

The excitability curve of Figure 10–22 is that of a large myelinated nerve fiber. The least possible voltage at which it will fire is called the *rheobase*, and the time required for this least voltage to stimulate the fiber is called the *utilization time* Then, if the voltage is increased to twice the rheobase voltage, the time required to stimulate the fiber is called the *chronaxie* The chronaxie is often used as a means of expressing relative excitabilities of different excitable tissues For instance, the chronaxie of a large type A fiber is about 0 0001 to 0.0002 second; of smaller myelinated nerve fibers, approximately 0.0003 second, of unmyelinated fibers, 0.0005 second; of skeletal muscle fibers, 0.00025 to 0 001 second; and of heart muscle, 0 001 to 0 003 second

The Refractory Period. A second action potential cannot occur in an excitable fiber as long as the membrane is still depolarized from the preceding action potential. The reason for this is that shortly after the action potential is initiated the sodium channels become inactivated, and any amount of excitatory signal applied to these channels at this point will not open the inactivation gates The only condition that will reopen them is for the membrane potential to return either to or almost to the original resting membrane potential level Then, within another small fraction of a second, the inactivation gates of the sodium channels open, and a new action potential can then be initiated

The period of time during which a second action potential cannot be elicited, even with a very strong stimulus, is called the *absolute refractory period* This period for large myelinated nerve fibers is about 1/2500 second. Therefore, one can readily calculate that such a fiber can carry a maximum of about 2500 impulses per second.

Following the absolute refractory period is a *relative refractory period* lasting about one quarter to one half as long. During this time stronger than normal stimuli can excite the fiber. The cause of this relative refractoriness is twofold. (1) During this time some of the sodium channels still have not been reversed from their inactivation state, and (2) the potassium channels are usually wide open at this time causing a state of hyperpolarization that makes it more difficult to stimulate the fiber.

INHIBITION OF EXCITABILITY—"STABILIZERS" AND LOCAL ANESTHETICS

Contrary to the factors that increase nerve excitability, still others, called *membrane-stabilizing factors*, can decrease excitability For instance, a *high extracellular fluid calcium ion concentration* decreases the membrane permeability and simultaneously reduces its excitability, as was explained earlier. Therefore, calcium ions are said to be a "stabilizer." Also, *low potassium ion concentration* in the extracellular fluids, because it has the direct effect of decreasing the permeability of the potassium channels, likewise acts as a stabilizer and reduces membrane excitability. Indeed, in a hereditary disease known as *familial periodic paralysis*, the extracellular potassium ion concentration is often so greatly reduced that the person actually becomes paralyzed but reverts to normal instantly after intravenous administration of potassium

Local Anesthetics. Among the most important stabilizers are the many substances used clinically as local anesthetics, including *cocaine, procaine, tetracaine*, and many other drugs. These act directly on the activation gates of the sodium channels, making it much more difficult for these gates to open and thereby reducing the membrane excitability. When the excitability has been reduced so low that the ratio of *action potential strength to excitability threshold* (called the "safety factor") is reduced below 1.0, a nerve impulse fails to pass through the anesthetized area

RECORDING MEMBRANE POTENTIALS AND ACTION POTENTIALS

The Cathode Ray Oscilloscope. Earlier in this chapter we noted that the membrane potential changes occur very rapidly throughout the course of an action potential Indeed, most of the action potential complex of large nerve fibers takes place in less than $\frac{1}{1000}$ second. In some figures of this chapter an electrical meter has been shown recording these potential changes However, it must be understood that any meter capable of recording them must be capable of responding extremely rapidly For practical purposes the only type of meter that is capable of responding accurately to the very rapid membrane potential changes of most excitable fibers is the cathode ray oscilloscope

Figure 10–23 illustrates the basic components of a cathode ray oscilloscope The cathode ray tube itself is composed basically of an *electron gun* and a *fluorescent surface* against which electrons are fired Where the electrons hit the surface, the fluorescent material glows If the electron beam is moved across the surface, the spot of glowing light also moves and draws a fluorescent line on the screen

In addition to the electron gun and fluorescent surface, the cathode ray tube is provided with two sets of plates:

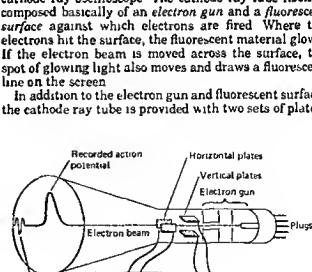


Figure 10–23. The cathode ray oscilloscope for recording transient action potentials

one set, called the *horizontal deflection plates*, is positioned on either side of the electron beam, and the other set, called the *vertical deflection plates*, is positioned above and below the beam If a negative charge is applied to the left-hand plate and a positive charge to the right-hand plate, the electron beam will be repelled away from the left plate and attracted toward the right plate, thus bending the beam toward the right, and this will cause the spot of light on the fluorescent surface of the cathode ray screen to move to the right Likewise, positive and negative charges can be applied to the vertical deflection plates to move the beam up or down

Since electrons travel at extremely rapid velocity and since the plates of the cathode ray tube can be alternately charged positively or negatively within less than a millionth of a second, it is obvious that the spot of light on the face of the tube can also be moved to almost any position in less than a millionth of a second. For this reason, the cathode ray tube oscilloscope can be considered to be an inertialess meter capable of recording with extreme fidelity almost any change in membrane potential

To use the cathode ray tube for recording action potentials, two electrical circuits must be employed These are (1) an *electronic sweep circuit* that controls the voltages on the horizontal deflection plates and (2) an *electronic amplifier* that controls the voltages on the vertical deflection plates. The sweep circuit automatically causes the spot of light to begin at the left-hand side and move slowly toward the right. When the spot reaches the right side it jumps back immediately to the left-hand side and starts a new trace.

The electronic amplifier amplifies signals that come from the nerve. If a change in membrane potential occurs while the spot of light is moving across the screen, this change in potential will be amplified and will cause the spot to rise above or fall below the mean level of the trace, as illustrated in the figure In other words, the sweep circuit provides the lateral movement of the electron beam while the amplifier provides the vertical movement in direct proportion to the changes in membrane potentials picked up by appropriate electrodes.

Figure 10–23 also shows that an electric stimulator is used to stimulate the nerve When the nerve is stimulated, a small *stimulus artifact* usually appears on the oscilloscope screen prior to the action potential

Recording the Monophasic Action Potential. Throughout this chapter "monophasic" action potentials have been shown in the different diagrams. To record these, an electrode such as that illustrated earlier in the chapter in Figure 10–3 must be inserted into the interior of the fiber Then, as the action potential spreads down the fiber, the changes in the potential inside the fiber are recorded as illustrated earlier in the chapter in Figures 10–7, 10–11, and 10–15

Recording a Biphasic Action Potential. When one wishes to record impulses from a whole nerve trunk, it is not feasible to place electrodes inside the nerve fibers Therefore, the usual method of recording is to place two electrodes on the outside of fibers However, the record that is obtained is then biphasic for the following reasons When an action potential moving down the nerve fiber reaches the first electrode, it becomes charged negatively while the second electrode is still unaffected This causes the oscilloscope to record in the negative direction Then as the action potential proceeds still farther down the nerve, there comes a point when the

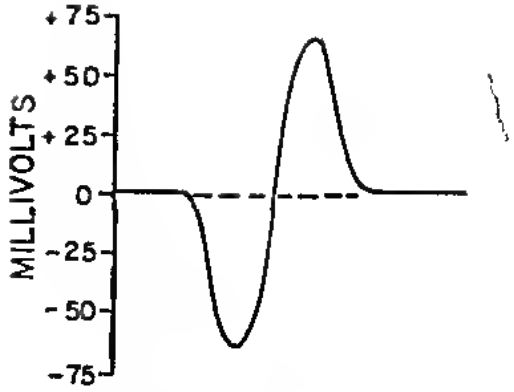


Figure 10–24. Recording of a biphasic action potential

membrane beneath the first electrode becomes repolarized while the second electrode is negative, and the oscilloscope records in the opposite direction. When these changes are recorded by the oscilloscope, a graphic record such as that illustrated in Figure 10–24 is recorded, showing a potential change first in one direction and then in the opposite direction.

REFERENCES

Adams, D J, *et al* Ionic currents in molluscan soma *Annu Rev Neurosci*, *3* 141, 1980

Adrian R H Charge movement in the membrane of striated muscle *Annu Rev Biophys Bioeng* 7 85 1978

Agnew W S Voltage-regulated sodium channel molecules *Annu Rev Physiol* *45* 517 1984

Aguayo, A J, and Karpati G (eds) Current Topics in Nerve and Muscle Research New York, Elsevier/North Holland, 1979

Armstrong C M Sodium channels and gating currents *Physiol Rev* *61* 644, 1981

Asbury, A. K., and Johnson P C Pathology of Peripheral Nerve Philadelphia, W B Saunders Co, 1978

Auerbach, A, and Sachs, F Patch clamp studies of single ionic channels *Annu Rev Biophys Bioeng*, *13* 269 1984

Baker, P F, *et al* Replacement of the axoplasm of giant nerve fibers with artificial solutions *J Physiol (Lond)*, *164* 330, 1962

Baker, P F, *et al* The effects of changes in internal ionic concentrations on the electrical properties of perfused giant axons. *J Physiol (Lond)* *164* 355 1962

Brinley, F J, Jr Calcium buffering in squid axons *Annu Rev Biophys Bioeng* 7 363, 1978

Carmeliet, E, and Vereecke, J Electrogenesis of the action potential and automaticity *In* Berne, R. M, *et al.* (eds) Handbook of Physiology Sec 2, Vol I Baltimore, Williams & Wilkins, 1979, p 269

Catterall, W A Neurotoxins that act on voltage-sensitive sodium channels in excitable membranes *Annu Rev Pharmacol Toxicol* *20* 15, 1980

Ceccarelli, B, and Clementi, F (eds) Neurotoxins, Tools in Neurobiology New York, Raven Press, 1979

Cole, K S Electrodiffusion models for the membrane of squid giant axon. *Physiol Rev*, *45* 340 1965

Cooper, S A New peripherally acting oral analgesic agents *Annu Rev Pharmacol Toxicol*, *23* 617, 1983

Cuénod, M, *et al* (eds.) Development and Chemical Specificity of Neurons New York, Elsevier North Holland, 1979

DiFrancesco D, and Noble, D A model of cardiac electrical activity incorporating ionic pumps and concentration changes *Phil Trans R Soc Lond (Biol)*, *307* 353, 1985

DiPolo, R., and Beaugé, L The calcium pump and sodium calcium exchange in squid axons *Annu Rev Physiol*, *45* 313, 1983

Fozzard, H A Conduction of the action potential *In* Berne, R. M, *et al* (eds) Handbook of Physiology Sec 2, Vol I Baltimore, Williams & Wilkins, 1979, p 335

French, R J, and Horn, R. Sodium channel gating: Models, mimics, and modifiers *Annu Rev Biophys Bioeng*, *12* 319, 1983

Goldin, S M, *et al.* Isolation and reconstitution of neuronal ion transport proteins *Annu Rev Neurosci*, *6* 419 1983

Gospodarowicz, D Epidermal and nerve growth factors in mammalian development *Annu Rev Physiol*, *43* 251, 1981

Greene, L. A , and Shooter, E. M . The nerve growth factor Biochemistry, synthesis, and mechanism of action *Annu Rev Neurosci* , *3* 353, 1980.
Guroff, G : Molecular Neurobiology New York, Marcel Dekkee, 1979
Harris, W. A • Neural activity and development *Annu Rev Physiol* , *43* 689, 1981.
Hille, B Gating in sodium channels of nerve *Annu Rev Physiol* , *38* 139, 1976
Hodgkin, A. L : The Conduction of the Nervous Impulse Springfield, Ill , Charles G Thomas, 1963
Hodgkin, A L , and Horowicz, P The effect of sudden changes in ionic concentrations on the membrane potential of single muscle fibers *J Physiol (Lond), 153* 370, 1960
Hodgkin, A L , and Huxley, A. F Movement of sodium and potassium ions during nervous activity *Cold Spr Harb Symp Quant Biol* , *17* 43, 1952
Hodgkin, A. L., and Huxley, A F Quantitative description of membrane current and its application to conduction and excitation in nerve *J Physiol (Lond), 117* 500, 1952
Jansson, G , Chemical neurotoxins as denervation tools in neurobiology *Annu Rev Neurosci* , *3* 169, 1980
Katz, B Nerve, Muscle, and Synapse New York, McGraw Hill, 1966
Keynes, R D Ion channels in the nerve cell membrane *Sci Am* *240*(3) 126, 1979
Kostyuk, P G Intracellular perfusion of nerve cells and its effects on membrane currents *Physiol Rev* , *64* 435, 1984
Latorre, R., and Alvarez, O.. Voltage-dependent channels in lipid bilayer membranes *Physiol Rev* , *61*,77, 1981
Latorre, R., *et al* K• channels gated by voltage and ions *Annu Rev Physiol* , *46* 485, 1984
Malhotra, S K The Plasma Membrane New York, John Wiley & Sons, 1983
Menaker, M , *et al* , The physiology of circadian pacemakers *Annu Rev Physiol* , *40* 501, 1978
Moody, W , Jr • Effects of intracellular H• on the electrical properties of excitable cells. *Annu Rev Neurosci* , *7* 257, 1984
Naftalin, R. J • The thermostatics and thermodynamics of cotransport *Biochim Biophys Acta, 778* 155, 1984
Noble, D . Applications of Hodgkin Huxley equations to excitable tissues *Physiol , Rev 46* 1, 1966
Patrick, J , *et al* Biology of cultured nerve and muscle *Annu Rev Neurosci* , *1* 417, 1978
Pfenninger, K H. Organization of neuronal membrane *Annu Rev Neurosci* , *1* 445, 1978
Rall, W Core conductor theory and cable properties of neurons *In* Brookhart, J M , and Mountcastle, V B (eds) Handbook of Physiology Sec 1, Vol 1 Baltimore, Williams & Wilkins, 1977, p 39
Requena, J Calcium transport and regulation in nerve fibers *Annu Rev. Biophys Bioeng* , *12* 237, 1983
Rogart, R Sodium channels in nerve and muscle membrane *Annu Rev Physiol* , *43* 711, 1981
Sakmann, B , and Neher, E • Patch clamp techniques for studying ionic channels in excitable membranes *Annu Rev Physiol* , *46* 455, 1984
Schubert, D Developmental Biology of Cultured Nerve, Muscle and Glia. New York, John Wiley & Sons, 1984
Schwartz, J H Axonal transport. components, mechanisms, and specificity *Annu Rev Neurosci* , *2* 476, 1979
Schwartz, W , and Passow, H Ca²⁺-activated K• channels in erythrocytes and excitable cells *Annu Rev Physiol* , *45* 359, 1983
Shanes, A M Electrochemical aspects of physiological and pharmacological action in excitable cells *Pharmacol Rev* , *10* 59, 165, 1958
Spodin, R A Ion Transport in Skeletal Muscle New York, John Wiley & Sons, 1982
Snell, R M. (ed) Transcellular Membrane Potentials and Ionic Fluxes New York, Gordon Press Pubs., 1984
Sperelakis, N Hormonal and neurotransmitter regulation of Ca•• influx through voltage-dependent slow channels in cardiac muscle membrane *Membr Biochem* 5 131, 1984
Stefani, E , and Chiarandini, D J Ionic channels in skeletal muscle *Annu Rev Physiol* , *44* 357, 1982
Stevens, C F The neuron *Sci Am* , *241*(3) 54, 1979
Swadlow, H A , *et al* Modulation of impulse conduction along the axonal tree *Annu Rev Biophys Bioeng* , *9* 143, 1980
Tsien, R W Calcium channels in excitable cell membranes *Annu. Rev Physiol* , *45* 341, 1983
Ulbricht W Kinetics of drug action and equilibrium results at the node of Ranvier *Physiol Rev* , *61* 785, 1981
Vinores, S. and Guroff G Nerve growth factar Mechanism of action *Annu Rev Biophys Bioeng* , *9* 223, 1980
Walker, J L , and Brown, H M Intracellular ionic activity measurements in nerve and muscle *Physiol Rev* , *57* 729, 1977
Wallick, E T , *et al* Biochemical mechanism of the sodium pump *Annu Rev Physiol* , *41* 397, 1979
Waxman, S G (ed) Physiology and Pathobiology of Axons New York, Raven Press 1978
Weiss, D C (ed) Axoplasmic Transport in Physiology and Pathology New York, Springer Verlag, 1982
Windhager, E E., and Taylor A Regulatory role of intracellular calcium ions in epithelial Na transport *Annu Rev Physiol* , *45* 519, 1983
Wright E M Electrophysiology of plasma membrane vesicles *Am J Physiol* , *246* F363, 1984
Zigmond, R E , and Bowers, C W Influence of nerve activity on the macromolecular content of neurons and their effector organs *Annu Rev Physiol* , *43* 673, 1981

11

Contraction of Skeletal Muscle

Approximately 40 per cent of the body is skeletal muscle and almost another 10 per cent is smooth and cardiac muscle. Many of the same principles of contraction apply to all these different types of muscle, but in the present chapter the function of skeletal muscle is considered mainly; the specialized functions of smooth muscle will be discussed in the following chapter, and cardiac muscle in Chapter 13.

PHYSIOLOGIC ANATOMY OF SKELETAL MUSCLE

THE SKELETAL MUSCLE FIBER

Figure 11–1 illustrates the organization of skeletal muscle, showing that all skeletal muscles are made of numerous fibers ranging between 10 and 80 microns in diameter. Each of these fibers in

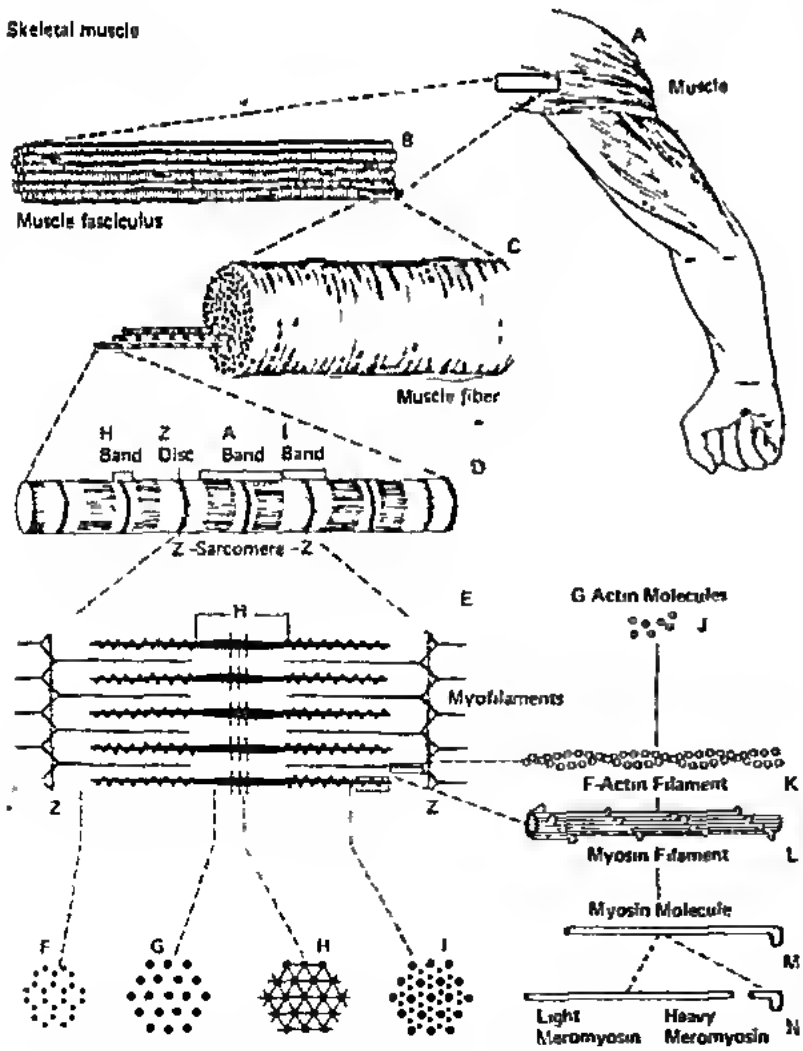


Figure 11–1. Organization of skeletal muscle, from the gross to the molecular level. *F, G, H,* and *I* are cross sections at the levels indicated. (Drawing by Sylvia Colard Keene. From Bloom and Fawcett. A Textbook of Histology. Philadelphia, W. B. Saunders Company, 1975.)

turn is made up of successively smaller subunits, also illustrated in Figure 11–1, that will be described in subsequent paragraphs:

In most muscles the fibers extend the entire length of the muscle, and, except for about 2 per cent of the fibers, each is innervated by only one nerve ending, located near the middle of the fiber.

The Sarcolemma. The sarcolemma is the cell membrane of the muscle fiber. However, the sarcolemma consists of a true cell membrane, called the *plasma membrane*, and an outer coat consisting of a thin layer of polysaccharide material containing numerous thin collagen fibrillae. At the end of the muscle fiber, this surface layer of the sarcolemma fuses with a tendon fiber, and the tendon fibers in turn collect into bundles to form the muscle tendons and thence insert into the bones.

Myofibrils; Actin and Myosin Filaments. Each muscle fiber contains several hundred to several thousand *myofibrils*, which are illustrated by the many small open dots in the cross-sectional view of Figure 11–1C. Each myofibril (Figure 11–1D) in turn has, lying side-by-side, about 1500 *myosin filaments* and 3000 *actin filaments*, which are large polymerized protein molecules that are responsible for muscle contraction. These can be seen in longitudinal view in the electron micrograph of Figure 11–2, and are represented diagrammatically in Figure 11–1E. The thick filaments are *myosin* and the thin filaments are *actin*. Note that the myosin and actin filaments partially interdigitate and thus cause the myofibrils to have alternate light and dark bands. The light bands, which contain only actin filaments, are called *I bands* because they are mainly *isotropic* to polarized light. The dark bands, which contain the myosin filaments as well as the end of the actin filaments where they overlap the myosin, are called *A bands* because they are *anisotropic* to polarized light. Note also the small projections from the sides of the myosin filaments. These are called *cross-bridges*. They protrude from the surfaces of the myosin filaments along the entire extent of the filament, except in the very center. It is interaction between these cross-bridges and the actin filaments that causes contraction.

Figure 11–1E also shows that the actin filaments are attached to the so-called *Z disc*, and the filaments extend on either side of the Z disc to interdigitate with the myosin filaments. The Z disc, which itself is composed of several filamentous proteins different from the actin and myosin filaments, also passes from myofibril to myofibril, attaching the myofibrils to each other all the way across the muscle fiber. Therefore, the entire muscle fiber has light and dark bands, as is also true of the individual myofibrils. These bands give skeletal and cardiac muscle their striated appearance.

Figure 11–2. Electron micrograph of muscle myofibrils, showing the detailed organization of actin and myosin filaments. Note the mitochondria lying between the myofibrils. (From Fawcett, The Cell. Philadelphia, W. B. Saunders Company, 1981.)

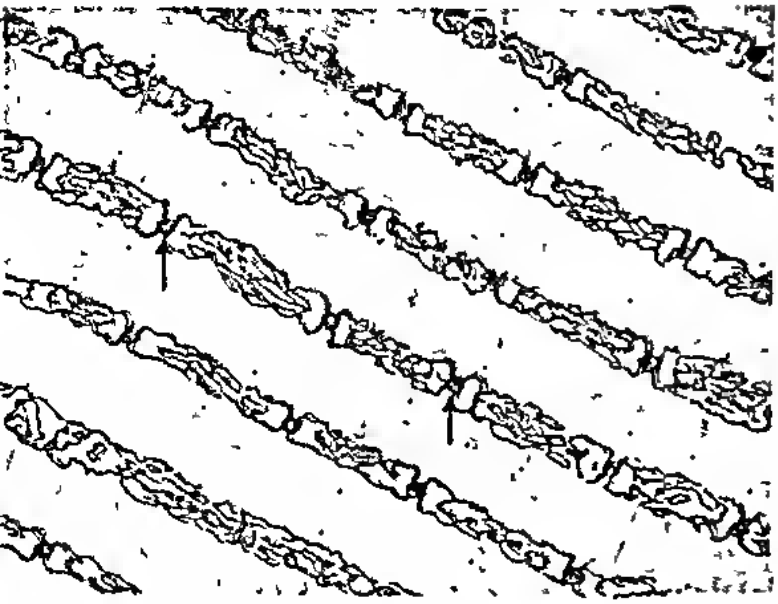


Figure 11–3. Sarcoplasmic reticulum surrounding the myofibril, showing the longitudinal system paralleling the myofibrils. Also shown in cross-section are the T tubules that lead to the exterior of the fiber membrane and that contain extracellular fluid (arrows). (from Fawcett. The Cell. Philadelphia, W. B. Saunders Company, 1981.)

The portion of a myofibril (or of the whole muscle fiber) that lies between two successive Z discs is called a *sarcomere*. When the muscle fiber is at its normal, fully stretched resting length, the length of the sarcomere is about 2.0 microns. At this length, the actin filaments completely overlap the myosin filaments and are just beginning to overlap each other. We shall see later that at this length the sarcomere also is capable of generating its greatest force of contraction.

When a muscle fiber is stretched beyond its resting length, as it is in Figure 11–1, the ends of the actin filaments pull apart, leaving a light area in the center of the A band. This light area, called the *H zone*, is illustrated in Figure 11–2. Such an H zone rarely occurs in the normally functioning muscle because normal sarcomere contraction occurs when the length of the sarcomere is between 2.0 microns and 1.6 microns. In this range the ends of the actin filaments not only overlap the myosin filaments but also overlap each other.

The Sarcoplasm. The myofibrils are suspended inside the muscle fiber in a matrix called *sarcoplasm*, which is composed of usual intracellular constituents. The fluid of the sarcoplasm contains large quantities of potassium, magnesium, phosphate, and protein enzymes. Also present are tremendous numbers of *mitochondria* that lie between and parallel to the myofibrils, a condition which is indicative of the great need of the contracting myofibrils for large amounts of ATP formed by the mitochondria.

The Sarcoplasmic Reticulum. Also in the sarcoplasm is an extensive endoplasmic reticulum, which in the muscle fiber is called the *sarcoplasmic reticulum*. This reticulum has a special organization that is extremely important in the control of muscle contraction, which will be discussed later in the chapter. The electron micrograph of Figure 11–3 illustrates the arrangement of this sarcoplasmic reticulum and shows how extensive it can be. The more rapidly contracting types of muscle have especially extensive sarcoplasmic reticula, indicating that this structure is important in causing rapid muscle contraction, as will also be discussed later.

MOLECULAR MECHANISM OF MUSCLE CONTRACTION

Sliding Mechanism of Contraction. Figure 11–4 illustrates the basic mechanism of muscle contraction. It shows the relaxed state of a sarcomere (above) and the contracted state (below). In the relaxed state, the ends of the actin filaments derived from two successive Z discs barely overlap each other while at the same time completely overlapping the myosin filaments. On the other hand, in the contracted state these actin filaments

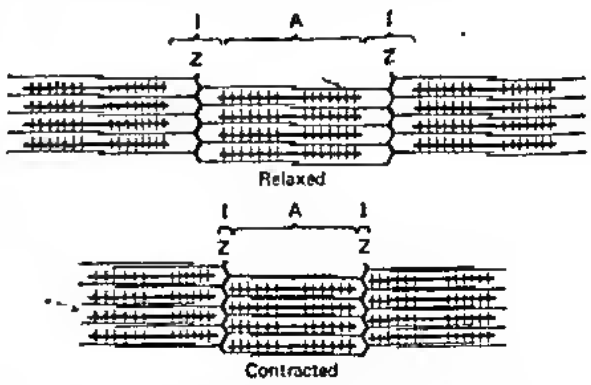


Figure 11–4. The relaxed and contracted states of a myofibril showing sliding of the actin filaments (black) into the channels between the myosin filaments (red).

have been pulled inward among the myosin filaments so that they now overlap each other to a major extent. Also, the Z discs have been pulled by the actin filaments up to the ends of the myosin filaments. Indeed, the actin filaments can be pulled together so tightly that the ends of the myosin filaments actually buckle during very intense contraction. Thus, muscle contraction occurs by a *sliding filament mechanism*.

But what causes the actin filaments to slide inward among the myosin filaments? Almost certainly, this is caused by mechanical, chemical, or electrostatic forces generated by the interaction of the cross-bridges of the myosin filaments with the actin filaments, as we shall discuss in the following sections.

Under resting conditions, the sliding forces between the actin and myosin filaments are inhibited, but when an action potential travels over the muscle fiber membrane, this causes the release of large quantities of calcium ions into the sarcoplasm surrounding the myofibrils. These calcium ions activate the forces between the filaments and contraction begins. But energy is also needed for the contractile process to proceed. This energy is derived from the high energy bonds of adenosine triphosphate (ATP), which is degraded to adenosine diphosphate (ADP) to give the energy required.

In the next few sections we will describe what is known about the details of the molecular processes of contraction. To begin this discussion, however, we must first characterize in detail the myosin and actin filaments.

MOLECULAR CHARACTERISTICS OF THE CONTRACTILE FILAMENTS

The Myosin Filament. The myosin filament is composed of multiple myosin molecules, each having a molecular weight of about 480,000. Figure 11–5A illustrates an individual molecule; section B illustrates the organization of the molecules to form a myosin filament, as well as its interaction with the ends of two actin filaments.

The *myosin molecule* is comprised of six polypeptide chains, two *heavy chains* each with a molecular weight of about 200,000 and four *light chains* with molecular weights of about 20,000 each. The two heavy chains coil around each other to form a double helix. However, one end of each of these chains is folded into a globular protein mass called the myosin *head*. Thus, there are two free heads lying side by side at one end of the double helix myosin molecule; the other end of the coiled helix is called the *tail*. The four light chains are also parts of the myosin heads, two to each head. These light chains help control the function of the head during the process of muscle contraction.

The *myosin filament* is made up of about 200 individual myosin molecules. The central portion

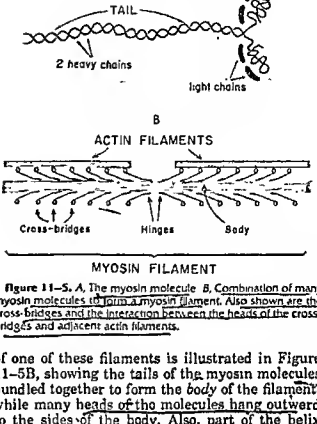


Figure 11–5. *A*, The myosin molecule. *B*, Combination of many myosin molecules to form a myosin filament. Also shown are the cross-bridges and the interaction between the heads of the cross-bridges and adjacent actin filaments.

of one of these filaments is illustrated in Figure 11–5B, showing the tails of the myosin molecules bundled together to form the *body* of the filament, while many heads of the molecules hang outward to the sides of the body. Also, part of the helix portion of each myosin molecule extends to the side along with the head, thus providing an *arm* that extends the head outward from the body as shown in the figure. The protruding arms and heads together are called *cross-bridges*, and each of these is believed to be flexible at two points called *hinges*, one where the arm leaves the body of the myosin filament and the other where the two heads attach to the arm. The hinged arms allow the heads to be extended either far outward from the body of the myosin filament or to be brought close to the body. The hinged heads are believed to participate in the actual contraction process, as we shall discuss in the following sections.

The total length of the myosin filament is 1.6 microns. However, note that there are no cross-bridge heads in the very center of the myosin filament for a distance of about 0.2 micron because the hinged arms extend toward both ends of the myosin filament away from the center; therefore, in the center there are only tails of the myosin molecules and no heads.

Now, to complete the picture, the myosin filament itself is twisted so that each successive set of cross-bridges is axially displaced from the previous set by 120 degrees. This insures that the cross-bridges extend in all directions around the filament.

ATPase Activity of the Myosin Head. Another feature of the myosin head that is essential for muscle contraction is that it can function as an

ATPase enzyme. As we shall see later, this property allows the head to cleave ATP and to use the energy derived from the ATP's high energy phosphate bond to energize the contraction process.

The Actin Filament. The actin filament is also complex. It is composed of three different components. *actin*, *tropomyosin*, and *troponin*

The backbone of the actin filament is a double-stranded F-actin protein molecule, illustrated in Figure 11–6 The two strands are wound in a helix in the same manner as the myosin molecule, but with a complete revolution every 70 nanometers.

Each strand of the double F-actin helix is composed of polymerized G-actin molecules, each having a molecular weight of 42,000 There are approximately 13 of these molecules in each revolution of each strand of helix Attached to each one of the G-actin molecules is one molecule of ADP. It is believed that these ADP molecules are the active sites on the actin filaments with which the cross-bridges of the myosin filaments interact to cause muscle contraction. The active sites on the two F-actin strands of the double helix are staggered, giving one active site on the overall actin filament approximately every 2 7 nanometers.

Each actin filament is approximately 1 micron long The bases of the actin filaments are inserted strongly into the Z discs, while their other ends protrude in both directions into the adjacent sarcomeres to lie in the spaces between the myosin molecules, as illustrated in Figure 11–4.

The Tropamyosin Strands. The actin filament also contains two additional protein strands that are polymers of *tropomyosin* molecules, each molecule having a molecular weight of 70,000 and extending a length of 40 nanometers It is believed that each tropomyosin strand is loosely attached to an F-actin strand and that in the resting state it physically covers the active sites of the actin strands so that interaction cannot occur between the actin and myosin to cause contraction.

Troponin and Its Role in Muscle Contraction. Attached approximately two thirds the distance along each tropomyosin molecule is a complex of three globular protein molecules called *troponin*. One of the globular proteins (troponin I) has a strong affinity for actin, another (troponin T) for tropomyosin, and a third (troponin C) for calcium ions This complex is believed to attach the tropomyosin to the actin. The strong affinity of the troponin for calcium ions is believed to initiate the contraction process, as will be explained in the following section.

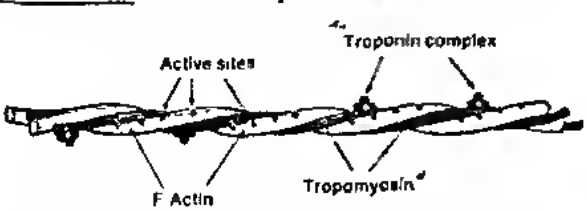


Figure 11–6. The actin filament composed of two helical strands of F-actin and two tropomyosin strands that lie in the grooves between the actin strands Attaching the tropomyosin to the actin are several troponin complexes

Interaction of Myosin and Actin Filaments to Cause Contraction

Inhibition of the Actin Filament by the Troponin-Tropomyosin Complex; Activation by Calcium Ions. A pure actin filament without the presence of the troponin-tropomyosin complex binds strongly with myosin molecules in the presence of magnesium ions and ATP, both of which are normally abundant in the myofibril. But, if the troponin-tropomyosin complex is added to the actin filament, this binding does not take place Therefore, it is believed that the active sites on the normal actin filament of the relaxed muscle are inhibited (or perhaps physically covered) by the troponin-tropomyosin complex. Consequently they cannot interact with the myosin filaments to cause contraction Before contraction can take place the inhibitory effect of the troponin tropomyosin complex must itself be inhibited.

Now, let us discuss the role of the calcium ions In the presence of large amounts of calcium ions the inhibitory effect of the troponin-tropomyosin on the actin filaments is itself inhibited The mechanism of this is not known, but one suggestion is the following When calcium ions combine with troponin C, each molecule of which can bind strongly with up to four calcium ions even when they are present in minute quantities, the troponin complex supposedly undergoes a conformational change that in some way tugs on the tropomyosin protein strand and supposedly moves the tropomyosin strand deeper into the groove between the two actin strands This "uncovers" the active sites of the actin, thus allowing contraction to proceed Though this is a hypothetical mechanism, nevertheless it does emphasize that the normal relationship between the tropomyosin-troponin complex and actin is altered by calcium ions—a condition that leads to contraction

Interaction Between the "Activated" Actin Filament and the Myosin Cross-Bridges—The "Walk-Along" Theory of Contraction. As soon as the actin filament becomes activated by the calcium ions, it is believed that the heads of the cross bridges from the myosin filaments immediately become attracted to the active sites of the actin filament, and this in some way causes contraction to occur. Though the precise manner by which this interaction between the cross-bridges and the actin causes contraction is still unknown a suggested hypothesis for which considerable evidence exists is the *"walk-along" theory of contraction.*

Figure 11–7 illustrates the postulated walk-along mechanism for contraction. This figure

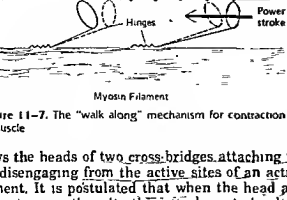


Figure 11–7. The "walk along" mechanism for contraction of the muscle

shows the heads of two cross-bridges attaching to and disengaging from the active sites of an actin filament. It is postulated that when the head attaches to an active site this attachment simultaneously causes profound changes in the intramolecular forces in the head and arm of the cross-bridge. The new alignment of forces causes the head to tilt toward the arm and to drag the actin filament along with it. This tilt of the head of the cross-bridge is called the *power stroke*. Then, immediately after tilting, the head automatically breaks away from the active site and returns to its normal perpendicular direction. In this position it combines with an active site farther down along the actin filament, then, a similar tilt takes place again to cause a new power stroke, and the actin filament moves another step. Thus, the heads of the cross-bridges bend back and forth and step by step walk along the actin filament, pulling the action toward the center of the myosin filament.

Each one of the cross-bridges is believed to operate independently of all others, each attaching and pulling in a continuous but rondom cycle. Therefore, the greater the number of cross-bridges in contact with the actin filament at any given time, the greater, theoretically, is the force of contraction.

ATP as the Source of Energy for Contraction—Chemical Events in the Motion of the Myosin Heads. When a muscle contracts against a load, work is performed and energy is required. It is found that large amounts of ATP are cleaved to form ADP during the contraction process. Furthermore, the greater the amount of work performed by the muscle, the greater the amount of ATP that is cleaved, which is called the *Fenn effect*. Though it is still not known exactly how ATP is used to provide the energy for contraction, the following is a sequence of events that has been suggested as the means by which this occurs:

1. Before contraction begins, the heads of the cross-bridges bind with ATP. The ATPase activity of the myosin head immediately cleaves the ATP but leaves the cleavage products, ADP plus Pi, bound to the head. In this state, the conformation of the head is such that it extends perpendicularly toward the actin filament but is not yet attached to the actin.

2. Next, when the inhibitory effect of the troponin-tropomyosin complex is itself inhibited by calcium ions, active sites on the actin filament are uncovered and the myosin heads do then bind with these, as illustrated in Figure 11–7.

3. The bond between the head of the cross-bridge and the active site of the actin filament causes a conformational change in the head, causing the head to tilt backward toward the arm of the cross-bridge. This provides the *power stroke* for pulling the actin filament. The energy that activates the power stroke is the energy already stored in the head at the time of cleavage of the ATP.

4. Once the head of the cross-bridge is tilted, this allows release of the ADP and Pi and exposes a site on the head where new ATP can bind. Therefore, a new molecule of ATP binds, and this binding in turn causes detachment of the head from the actin.

5. After the head has split away from the actin, the new molecule of ATP is also cleaved, and the energy again "cocks" the head back to its perpendicular condition ready to begin a new power stroke cycle.

6. Then, when the cocked head, with its stored energy derived from the cleaved ATP, binds with a new active site on the actin filament, it becomes uncocked and once again provides the power stroke.

7. Thus, the process proceeds again and again until the actin filament pulls the Z membrane up against the ends of the myosin filaments or until the load on the muscle becomes too great for further pulling to occur.

DEGREE OF ACTIN AND MYOSIN FILAMENT OVERLAP—EFFECT ON TENSION DEVELOPED BY THE CONTRACTING MUSCLE

Figure 11–8 illustrates the effect of sarcomere length on the tension developed by a single, contracting, *isolated* muscle fiber. To the right are

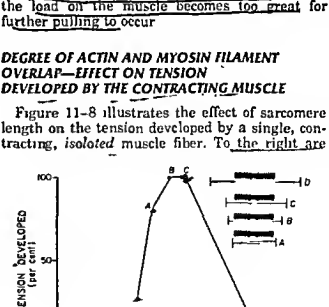


Figure 11–8. Length tension diagram for a single sarcomere, illustrating maximum strength of contraction when the sarcomere is 2.0 to 2.2 microns in length. At the upper right are shown the relative positions of the actin and myosin filaments at different sarcomere lengths from point A to point D. (Modified from Gordon, Huxley, and Julian: *J Physiol.* 171:287, 1964.)

illustrated different degrees of overlap of the myosin and actin filaments at different sarcomere lengths. At point D on the diagram, the actin filament has pulled all the way out to the end of the myosin filament with no overlap at all. At this point, the tension developed by the activated muscle is zero. Then, as the sarcomere shortens and the actin filament overlaps the myosin filament progressively more and more, the tension increases progressively until the sarcomere length decreases to about 2.2 microns. At this point the actin filament has already overlapped all the cross-bridges of the myosin filament but has not yet reached the center of the myosin filament. Upon further shortening, the sarcomere maintains full tension until point B at a sarcomere length of approximately 2.0 microns. At this point the ends of the two actin filaments begin to overlap. As the sarcomere length falls from 2 microns down to about 1.65 microns, at point A, the strength of contraction decreases. It is at this point that the two Z discs of the sarcomere abut the ends of the myosin filaments. Then, as contraction proceeds to still shorter sarcomere lengths, the ends of the myosin filaments are actually crumpled, and as illustrated in Figure 11–8 the strength of contraction also decreases precipitously.

This diagram illustrates that maximum contraction occurs when there is maximum overlap between the actin filaments and the cross-bridges of the myosin filaments, and it supports the idea that the greater the number of cross-bridges pulling the actin filaments, the greater the strength of contraction

Relation of Force of Contraction of the Intact Muscle to Muscle Length. The upper curve of Figure 11–9 is similar to that in Figure 11–8, but this illustrates the intact whole muscle rather than the isolated muscle fiber. The whole muscle has a large amount of connective tissue in it; also, the sarcomeres in different parts of the muscle do not necessarily contract exactly in unison. Therefore, the curve has somewhat different dimensions from those illustrated for the individual muscle fiber, but it nevertheless exhibits the same form.

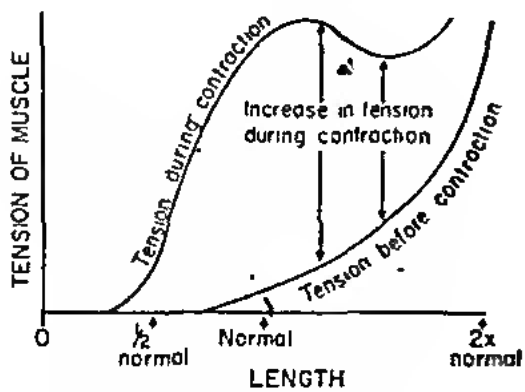


Figure 11–9. Relation of muscle length to force of contraction.

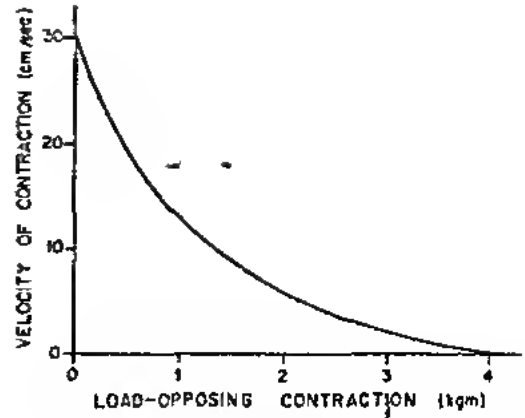


Figure 11–10. Relation of load to velocity of contraction in a skeletal muscle 8 centimeters long

Note in Figure 11–9 that when the muscle is at its normal resting length and is then activated, it contracts with maximum force of contraction. If the muscle is stretched to much greater than normal length prior to contraction, a large amount of *resting tension* develops in the muscle even before contraction takes place; this tension results from the elastic forces of the connective tissue, of the sarcolemma, the blood vessels, the nerves, and so forth. However, the *increase* in tension during contraction, called *active tension*, decreases as the muscle is stretched beyond its normal length. This is demonstrated by the decrease in tha arrow length in the figure.

RELATION OF VELOCITY OF CONTRACTION TO LOAD

A muscle contracts extremely rapidly when it contracts against no load—to a state of full contraction in approximately 0.1 second for the average muscle. However, when loads are applied, the velocity of contraction becomes progressively less as the load increases, as illustrated in Figure 11–10. When the load increases to equal the maximum force that the muscle can exert, then the velocity of contraction becomes zero and no contraction at all results, despite activation of the muscle fiber

This decreasing velocity with load is caused by the fact that a load on a contracting muscle is a reverse force that opposes the contractile force caused by muscle contraction Therefore, the net force that is available to cause velocity of shortening is correspondingly reduced

INITIATION OF MUSCLE CONTRACTION: EXCITATION-CONTRACTION COUPLING

Initiation of contraction in skeletal muscle begins with action potentials in the muscle fibers. These elicit electrical currents that spread to the interior of the fiber where they cause release of calcium ions from the sarcoplasmic reticulum. It is the calcium ions that in turn initiate the chem-

ical events of the contractile process. This overall process for controlling muscle contraction is called *excitation-contraction coupling*.

THE MUSCLE ACTION POTENTIAL

Almost everything discussed in Chapter 10 regarding initiation and conduction of action potentials in nerve fibers applies equally well to skeletal muscle fibers, except for quantitative differences. Some of the quantitative aspects of muscle potentials are the following:

1. Resting membrane potential: Approximately −90 millivolts in skeletal fibers—the same as in large myelinated nerve fibers.
2. Duration of action potential: 1 to 5 milliseconds in skeletal muscle—about five times as long as large myelinated nerves.
3. Velocity of conduction: 3 to 5 meters per second—about 1/13 the velocity of conduction in the large myelinated nerve fibers that excite skeletal muscle.

Excitation of Skeletal Muscle Fibers by Nerves. In normal function of the body, skeletal muscle fibers are excited by large myelinated nerve fibers. These attach to the skeletal muscle fibers at the neuromuscular junction, which will be discussed in detail in the following chapter. Except for 2 per cent of the muscle fibers, there is only one neuromuscular junction to each muscle fiber; this junction is located near the middle of the fiber. Therefore, the action potential spreads from the middle of the fiber toward its two ends. This dual direction of spreading from the center is important because it allows nearly coincident contraction of all sarcomeres of the muscles so that they can all contract together rather than separately.

Spread of the Action Potential to the Interior of the Muscle Fiber by Way of the Transverse Tubule System

The skeletal muscle fiber is so large that action potentials spreading along its surface membrane cause almost no current flow deep within the fiber. Yet, to cause contraction, these electrical currents must penetrate to the vicinity of all the separate myofibrils. This is achieved by transmission of the action potentials along *transverse tubules* (T tubules) that penetrate all the way through the muscle fiber from one side to the other. The T tubule action potentials in turn cause the sarcoplasmic reticulum to release calcium ions in the immediate vicinity of all the myofibrils, and it is these calcium ions that in turn cause contraction. Now, let us describe this system in much greater detail.

The Transverse Tubule–Sarcoplasmic Reticulum System. Figure 11–11 illustrates a group of myofibrils surrounded by the transverse tubule–sarcoplasmic reticulum system. The transverse tubules penetrate all the way from one side of the muscle fiber to the opposite side. Not shown in the figure is the fact that these tubules branch among themselves so that they form entire *planes* of T tubules interlacing among all the separate myofibrils. Also, it should be noted that where the T tubules originate from the cell membrane they are open to the exterior. Therefore, they communicate with the fluid surrounding the muscle fiber and contain extracellular fluid in their lumens. In other words, the T tubules are internal extensions of the cell membrane. Therefore, when an action potential spreads over a muscle fiber membrane, it spreads along the T tubules to the deep interior of the muscle fiber as well. The action potential currents surrounding these transverse tubules then elicit the muscle contraction.

Figure 11–11 shows the extensiveness of the *sarcoplasmic reticulum* as well. This is composed of two major parts: (1) long *longitudinal tubules* that terminate in (2) large chambers called *terminal cisternae* that abut the transverse tubules. When the muscle fiber is sectioned longitudinally and electron micrographs are made, one sees this abutting of the cisternae against the transverse tubule, which gives the appearance of a *triad* with a small central tubule and a large cisterna on either side. This is illustrated in Figure 11–11 and is also seen in the electron micrograph of Figure 11–3.

In the muscle of lower animals such as the frog, there is a single T tubule network for each sarcomere, located at the level of the Z disc as illustrated in Figure 11–11. Cardiac muscle also has this type of T tubule system. However, in mammalian skeletal muscle there are two T tubule networks for each sarcomere located near the two ends of the myosin filaments, which are the points where the actual mechanical forces of muscle contraction are created. Thus, mammalian skeletal muscle is optimally organized for rapid excitation of muscle contraction.

RELEASE OF CALCIUM IONS BY THE SARCOPLASMIC RETICULUM

One of the special features of the sarcoplasmic reticulum is that it contains calcium ions in very high concentration, and many of these ions are released when the adjacent T tubule is excited.

Figure 11–12 shows that the action potential of the T tubule causes current flow through the cisternae where they abut the T tubule. The cisternae project *junctional feet* that surround the T tubule, presumably facilitating passage of electrical current from the T tubule into the cisternae. This current flow causes rapid release of calcium ions from the cisternae and perhaps from the longitudinal tubules as well. Presumably this results from the opening of calcium channels similar to the

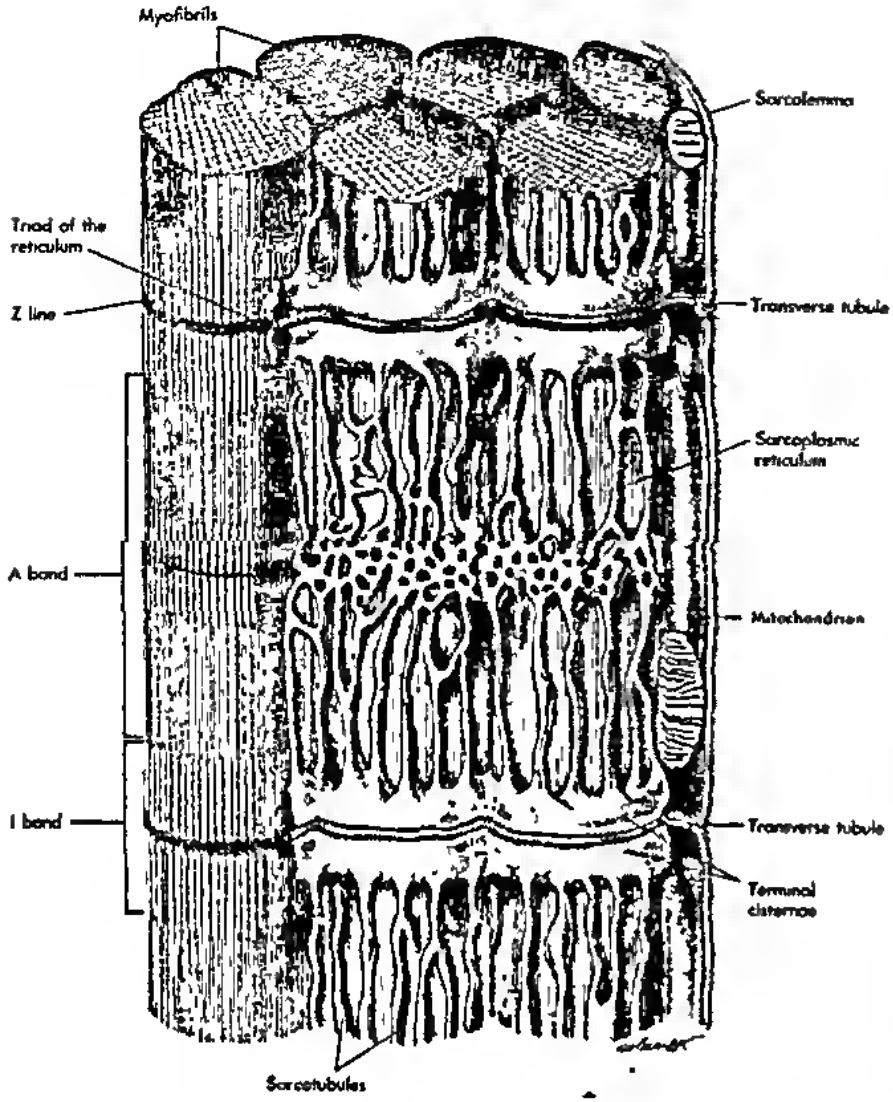


Figure 11–11. The transverse tubule sarcoplasmic reticulum system. Note the *longitudinal tubules* that terminate in large *cisternae*. The cisternae in turn abut the transverse tubules. Note also that the transverse tubules communicate with the outside of the cell membrane. This illustration was drawn from frog muscle, which has one transverse tubule per sarcomere, located at the Z line. A similar arrangement is found in mammalian heart muscle, but mammalian skeletal muscle has two transverse tubules per sarcomere, located at the A I junctions. (From Bloom and Fawcett: A Textbook of Histology. Philadelphia: W. B. Saunders Company, 1975. Modified after Peachey: *J. Cell Biol.* 25:209, 1965. Drawn by Sylvia Colard Keene.)

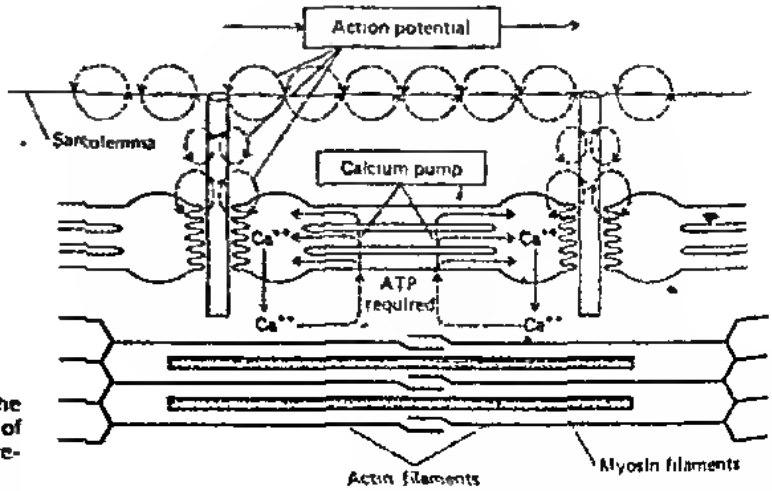


Figure 11–12. Excitation contraction coupling in the muscle, showing an action potential that causes release of calcium ions from the sarcoplasmic reticulum and then reuptake of the calcium ions by a calcium pump.

opening of sodium channels at the onset of the action potential, though the actual mechanism is still unknown.

The calcium ions that are thus released from the sarcoplasmic reticulum diffuse to the adjacent myofibrils where they bind strongly with troponin C_1 as discussed in an earlier section, and this in turn elicits the muscle contraction, as has also been discussed. However, the calcium ions also bind less strongly with the myosin filaments, which theoretically could be another factor in initiating contraction.

The Calcium Pump for Removing Calcium Ions from the Sarcoplasmic Fluid. Once the calcium ions have been released from the cisternae and have diffused to the myofibrils, muscle contraction will then continue as long as the calcium ions remain in high concentration in the sarcoplasmic fluid. However, a continually active calcium pump located in the walls of the sarcoplasmic reticulum pumps calcium ions out of the sarcoplasmic fluid back into the vesicular cavities of the reticulum. This pump can concentrate the calcium ions about 10,000-fold inside the sarcoplasmic reticulum. In addition, inside the reticulum is a protein called *calsequestrin* that can bind 40 times as much calcium as that in the ionic state, thus providing another 40-fold increase in the storage of calcium. Thus, this massive transfer of calcium in the sarcoplasmic reticulum causes total depletion of calcium ions in the fluid of the myofibrils. Therefore, except immediately after an action potential, the calcium ion concentration in the myofibrils is kept at an extremely low level.

The Excitatory "Pulse" of Calcium Ions. The normal concentration (less than 10^{-7} molar of calcium ions in the cytosol) that bathes the myofibrils is too little to elicit contraction. Therefore, in the resting state, the troponin-tropomyosin complex keeps the actin filaments inhibited and maintains a relaxed state of the muscle.

On the other hand, full excitation of the T tubule–sarcoplasmic reticulum system causes enough release of calcium ions to increase the concentration in the myofibrillar fluid to as high as 2×10^{-4} molar concentration, which is several times the level required to cause maximum muscle contraction (about 2×10^{-5} molar). Immediately thereafter, the calcium pump depletes the calcium ions again. The total duration of this calcium "pulse" in the usual skeletal muscle fiber lasts about 1/20 of a second, though it may last several times as long as this in some skeletal muscle fibers and be several times shorter in others (in heart muscle the pulse lasts for as long as 0.3 second because of the long duration of the cardiac action potential). It is during this calcium pulse that muscle contraction occurs. If the contraction is to continue without interruption for longer intervals, a series of such pulses must be initiated by a continuous series of repetitive action potentials, as will be discussed in more detail later in the chapter.

THE SOURCE OF ENERGY FOR MUSCLE CONTRACTION

We have already seen that muscle contraction depends upon energy supplied by ATP. Most of this energy is required to actuate the "walk-along" mechanism by which the cross-bridges pull the actin filaments, but small amounts are required for (1) pumping calcium from the sarcoplasm into the sarcoplasmic reticulum, and (2) pumping sodium and potassium ions through the muscle fiber membrane to maintain an appropriate ionic environment for the propagation of action potentials.

However, the concentration of ATP present in the muscle fiber, about 4 mM, is sufficient to maintain full contraction for only a few seconds at most. Fortunately, after the ATP is broken into ADP, as was described in Chapter 2, the ADP is rephosphorylated to form new ATP within a fraction of a second. There are several sources of the energy for this rephosphorylation:

The first source of energy that is used to reconstitute the ATP is the substance *phosphocreatine*, which carries a high energy phosphate bond similar to those of ATP. The high energy phosphate bond of the phosphocreatine has a slightly higher amount of free energy than that of the ATP bond. Therefore, it is instantly cleaved and the released energy causes bonding of a new phosphate ion to ADP to reconstitute the ATP. However, the total amount of phosphocreatine is also very little—only about five times as great as the ATP. Therefore, the combined energy of both the stored ATP and the phosphocreatine in the muscle is still capable of causing maximal muscle contraction for no longer than a few more seconds.

The next source of energy used to reconstitute both the phosphocreatine and the ATP is energy released from foodstuffs—from carbohydrates, fats, and proteins. A small amount of this energy is released during the initial breakdown of glucose and glycogen in the cells, which is the process of *glycolysis*. However, about 95 per cent of the energy is released during final oxidation of the foodstuffs, which occurs almost entirely in the mitochondria. Both these processes utilize the energy released from the foodstuffs to form new ATP. The importance of glycolysis is that energy can be released to form new ATP about 2.5 times as rapidly by this mechanism as by the oxidative mechanism. However, the glycolytic mechanism also rapidly builds up glycolytic end-products in the muscle cells so that glycolysis can usually sustain maximum muscle contraction for only about one minute. On the other hand, oxidative release of energy is exceedingly efficient and can also use other food substrates such as fats and

proteins in addition to glucose and glycogen, thus allowing continued muscle activity for many hours, though at a level of energy release only about one fourth that which can be achieved during the first few seconds of exercise.

The detailed mechanisms of these energetic processes are discussed in Chapters 67 through 71. In addition, the importance of the different mechanisms of energy release in different sports is discussed in Chapter 84 on sports physiology.

Efficiency of Muscle Contraction. The "efficiency" of an engine or a motor is calculated as the percentage of energy input that is converted into work instead of heat. The percentage of the input energy to a muscle (the chemical energy in the nutrients) that can be converted into work is less than 20 to 25 per cent, the remainder becoming heat. The reason for this low efficiency is that about half of the energy in the foodstuffs is lost during the formation of ATP, and even then only 40 to 45 per cent of the energy in the ATP itself can later be converted into work. Maximum efficiency can be realized only when the muscle contracts at a moderate velocity. If the muscle contracts very slowly or without any movement at all, large amounts of maintenance heat are released during the process of contraction even though little or no work is being performed, thereby decreasing the efficiency. On the other hand, if contraction is too rapid, large proportions of the energy are used to overcome the viscous friction within the muscle itself, and this, too, reduces the efficiency of contraction. Ordinarily, maximum efficiency is developed when the velocity of contraction is about 30 per cent of maximum.

CHARACTERISTICS OF A SINGLE MUSCLE TWITCH

Many features of muscle contraction can be especially well demonstrated by eliciting single *muscle twitches*. This can be accomplished by instantaneously exciting the nerve to a muscle or by passing a short electrical stimulus through the muscle itself, giving rise to a single, sudden contraction lasting for a fraction of a second.

Isometric Versus Isotonic Contraction. Muscle contraction is said to be *isometric* when the muscle does not shorten during contraction and *isotonic* when it shortens with the tension on the muscle remaining constant. Systems for recording the two types of muscle contraction are illustrated in Figure 11–13.

To the right is the isometric system in which the muscle is suspended between a solid rod and a lever of an electronic force transducer. This transducer records [illegible] with almost zero movement of the lever; therefore, [illegible] effect, the muscle is bound between fixed points so [illegible] it cannot contract significantly. To the left is shown an isotonic recording system: the muscle simply lifts a pan of weights so that the force against which the muscle contracts remains constant, though the length of the muscle changes considerably.

There are several basic differences between isometric and isotonic contractions. First, isometric contraction does not require much sliding of myofibrils among each other. Second, in isotonic contraction a load is moved, which involves the phenomenon of inertia. That is, the

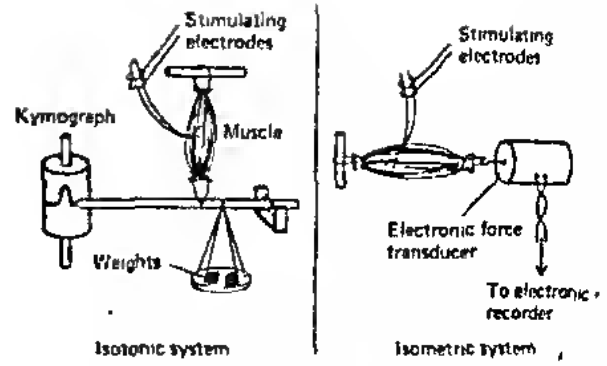


Figure 11–13. Isotonic and isometric recording systems.

weight or other type of object being moved must first be accelerated, and once a velocity has been attained the load has momentum that causes it to continue moving even after the contraction is over. Therefore an isotonic contraction is likely to last considerably longer than an isometric contraction of the same muscle. Third, isotonic contraction entails the performance of external work. Therefore, in accordance with the Fenn effect discussed previously, a greater amount of energy is used by the muscle.

In comparing the rapidity of contraction of different types of muscle, isometric recordings such as those illustrated in Figure 11–14 are usually used instead of isotonic recordings because the duration of an isotonic recording is almost as dependent on the inertia of the recording system as on the contraction itself, and this makes it difficult to compare time relationships of contractions from one muscle to another.

Muscles can contract both isometrically and isotonically in the body, but most contractions are actually a mixture of the two. When standing, a person tenses the quadriceps muscles to tighten the knee joints and to keep the legs stiff. This is isometric contraction. On the other hand, when a person lifts a weight using the biceps, this is mainly an isotonic contraction. Finally, contractions of leg muscles during running are a mixture of isometric and isotonic contractions—isometric mainly to keep the limbs stiff when the legs hit the ground and isotonic mainly to move the limbs.

The Series Elastic Component of Muscle Contraction. When muscle fibers contract against a load, those

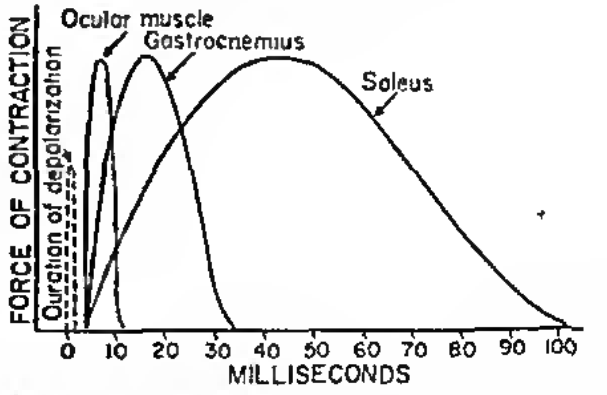


Figure 11–14. Duration of isometric contractions of different types of mammalian muscles, showing also a latent period between the action potential and muscle contraction.

portions of the muscle that do not contract—the tendons, the sarcolemmal ends of the muscle fibers where they attach to the tendons, and perhaps even the hinged arms of the cross-bridges—will stretch slightly as the tension increases. Consequently, the muscle must shorten an extra 3 to 5 per cent to make up for the stretch of these elements. The elements of the muscle that stretch during contraction are called the *series elastic component* of the muscle.

Characteristics of Isometric Twitches Recorded from Different Muscles. The body has many different sizes of skeletal muscles—from the very small stapedius muscle of only a few millimeters length and a millimeter or so in diameter up to the very large quadriceps muscle. Furthermore, the fibers may be as small as 10 microns in diameter or as large as 80 microns. And, finally, the energetics of muscle contraction vary considerably from one muscle to another. These different physical and chemical characteristics often manifest themselves in the form of different characteristics of contraction—some muscles contract rapidly while others contract slowly.

Figure 11-14 illustrates isometric contractions of three different types of skeletal muscles. an ocular muscle, which has a duration of contraction of less than 1/100 second; the gastrocnemius muscle, which has a duration of contraction of about 1/10 second, and the soleus muscle, which has a duration of contraction of about 1/5 second. It is interesting that these durations of contractions are adapted to the function of each of the respective muscles, for ocular movements must be extremely rapid to maintain fixation of the eyes upon specific objects, the gastrocnemius muscle must contract moderately rapidly to provide sufficient velocity of limb movement for running and jumping, while the soleus muscle is concerned principally with slow reactions for continual support of the body against gravity.

Fast Versus Slow Muscle Fibers. As we shall discuss more fully in Chapter 84 on sports physiology, every muscle of the body is composed of a mixture of so-called *fast* and *slow* muscle fibers, with still other fibers graduated between these two extremes. The muscles that react very rapidly are composed mainly of the fast fibers with only small numbers of the slow variety. And, conversely, the muscles that respond slowly but with a prolonged period of contraction are composed mainly of the slow type of fibers. The differences between these two types of fibers are the following:

Fast fibers (1) Much larger fibers for great strength of contraction. (2) Extensive sarcoplasmic reticulum for rapid release of calcium ions to initiate contraction. (3) Large amounts of glycolytic enzymes for rapid release of energy by the glycolytic process. (4) Less extensive blood supply because oxidative metabolism is of secondary importance. (5) Fewer mitochondria, also because oxidative metabolism is secondary

Slow fibers: (1) Smaller fibers. (2) Also innervated by smaller nerve fibers. (3) More extensive blood vessel system to supply extra amounts of oxygen. (4) Greatly increased numbers of mitochondria, also to support high levels of oxidative metabolism. (5) Fibers contain large amounts of myoglobin, an iron-containing protein similar to hemoglobin in red blood cells. Myoglobin combines with oxygen and holds this inside the muscle cell until it is needed by the mitochrondria.

From these descriptions, one can see that the fast fibers are adapted for very rapid and very powerful muscle contractions, such as for jumping or for short-distance powerful running. On the other hand, the slow fibers are adapted for prolonged, continued muscle activity, such as support of the body against gravity and long-continuing athletic events like marathon races.

The slow muscle is frequently also called *red muscle* because the excess myoglobin in the slow fibers gives them a slightly reddish tint. On the other hand, the relative deficiency of myoglobin in the fast fibers gives them a whitish appearance, so that this muscle is frequently called *white muscle*.

MECHANICS OF SKELETAL MUSCLE CONTRACTION

THE MOTOR UNIT

Each motor neuron that leaves the spinal cord usually innervates many different muscle fibers, the number depending on the type of muscle. All the muscle fibers innervated by a single motor nerve fiber are called a *motor unit*. In general, small muscles that react rapidly and whose control is exact have few muscle fibers (as few as two to three in some of the laryngeal muscles) in each motor unit and have a large number of nerve fibers going to each muscle. On the other hand, the large muscles that do not require a very fine degree of control, such as the gastrocnemius muscle, may have several hundred muscle fibers in a motor unit. An average figure for all the muscles of the body can be considered to be about 150 muscle fibers to the motor unit.

Usually muscle fibers of adjacent motor units overlap, with small bundles of 10 to 15 fibers from one motor unit lying among similar bundles of the second motor unit. This interdigitation allows the separate motor units to contract in support of each other rather than entirely as individual segments

Macromotor Units. Loss of some of the nerve fibers to a muscle causes the remaining nerve fibers to sprout forth and innervate many of the paralyzed muscle fibers. When this occurs, such as following poliomyelitis, one occasionally develops *macromotor units*, which can contain as many as five times the normal number of muscle fibers This obviously decreases the degree of control that one has over the muscles but, nevertheless, allows the muscles to regain function

total (addition)

SUMMATION OF MUSCLE CONTRACTION

Summation means the adding together of individual muscle twitches to make strong and concerted muscle movements. In general, summation occurs in two different ways. (1) by increasing the number of motor units contracting simultaneously, and (2) by increasing the rapidity of contraction of individual motor units. These are called, respectively, *multiple motor unit summation* and *wave summation* (or spatial summation and temporal summation).

Multiple Motor Unit Summation. Even within a single muscle, the numbers of muscle fibers and their sizes in the different motor units vary tremendously, so that one motor unit may be as much as 50 times as strong as another. The smaller motor units are far more easily excited than are the larger ones [illegible] y are innervated by sma[illegible] r nerve fibe[illegible]

the spinal cord have a naturally high level of excitability. This effect causes the gradations of muscle strength during weak muscle contraction to occur in very small steps, while the steps become progressively greater as the intensity of contraction increases because the larger motor units then begin to contract.

Wave Summation. Figure 11–15 illustrates the principles of wave summation, showing in the lower left-hand corner several single muscle twitches followed by successive muscle twitches at increasing frequencies. When the frequency of twitches rises above 10 per second, the first muscle twitch is not completely over by the time the second one begins. Therefore, since the muscle is already in a partially contracted state when the second twitch begins, the degree of muscle shortening this time is slightly greater than that which occurs with the single muscle twitch. At more rapid rates of contraction, the degree of summation of successive contractions becomes greater and greater, because the successive contractions appear at earlier times following the preceding contraction.

Tetanization. When a muscle is stimulated at progressively greater frequencies, a frequency is finally reached at which the successive contractions fuse together and cannot be distinguished one from the other. This state is called *tetanization*; and the lowest frequency at which it occurs is called the *critical frequency*.

Tetanization results partly from the viscous properties of the muscle and partly from the nature of the contractile process itself. The muscle fibers are filled with sarcoplasm, which is a viscous fluid, and the fibers are encased in fasciae and muscle sheaths that have a viscous resistance to change in length. Therefore, these viscous factors play a role in causing the successive contractions to fuse with each other.

But in addition to the viscous property of muscle, the activation process itself lasts for a definite period of time, and successive pulsatile states of activation of the muscle fiber can occur so rapidly that they fuse into a long continual state of activation, that is, the level of free calcium ions in the myofibrils remains continuously above the level required for full activation of the contractile process, providing an uninterrupted stimulus for maintenance of contraction. Once the critical frequency for tetanization is reached, further increase in rate of stimulation increases the force of contraction only a few more per cent, as shown in Figure 11–15.

Asynchronous Summation of Motor Units. Actually it is rare for either multiple motor unit summation or wave summation to occur separately from each other in normal muscle function. Instead, special neurogenic mechanisms in the spinal cord normally increase both the impulse rate and the number of motor units firing at the same time. If a motor unit fires at all, it usually fires at least five times per second, but this can increase to as high as 50 per second for most muscles or much more than this for the very fast muscles—to frequencies sufficient to cause complete tetanization.

Yet, even when tetanization of individual motor units of a muscle is not occurring, the tension exerted by the whole muscle is still continuous and nonjerky because *the different motor units fire asynchronously*, that is, while one is contracting another is relaxing; then another fires, followed by still another, and so forth. Consequently, even when motor units fire as infrequently as five times per second, the muscle contraction, though weak, is nevertheless very smooth.

Maximum Strength of Contraction. The maximum strength of tetanic contraction of a muscle operating at a normal muscle length is about 3.5 kilograms per square centimeter of muscle, or 50 pounds per square inch. Since a quadriceps muscle can at times have as much as 16 square inches of muscle belly, as much as 800 pounds of tension may at times be applied to the patellar tendon. One can readily understand, therefore, how it is possible for muscle sometimes to pull their tendons out of the insertions in bones. This often occurs where the patellar tendon inserts in the tibia, and it occurs even more frequently where the Achilles tendon of the gastrocnemius muscle inserts at the heel.

Changes in Muscle Strength at the Onset of Contraction—The Staircase Effect (Treppe). When a muscle begins to contract after a long period of rest, its initial strength of contraction may be as little as one-half its strength 30 to 50 muscle twitches later. That is, the strength of contraction increases to a plateau, a phenomenon called the *staircase effect* or *treppe*.

Though all the possible causes of the staircase effect are not yet known, it is believed to be caused primarily by electrolyte changes that occur when a series of contractions begins. For instance, there is a net increase in calcium ions inside the muscle fiber because of movement of calcium ions inward through the membrane with each action potential. There is probably also further increase of calcium ions in the cytosol because of release of these ions from the sarcoplasmic reticulum and failure to recapture the ions immediately. In addition, there is decreased potassium inside the cell as well as increased sodium. It has been suggested that the changes in these two ions increase the rate of liberation of calcium ions from the sarcoplasmic reticulum. Recalling the earlier discussion of the relationship of calcium ions to the contractile process, one can readily understand that progressive increase in calcium ion concentration in the sarcoplasm, caused either directly or as a consequence of sodium and potassium movement, could progressively increase the strength of muscle contraction, giving rise to the staircase effect.

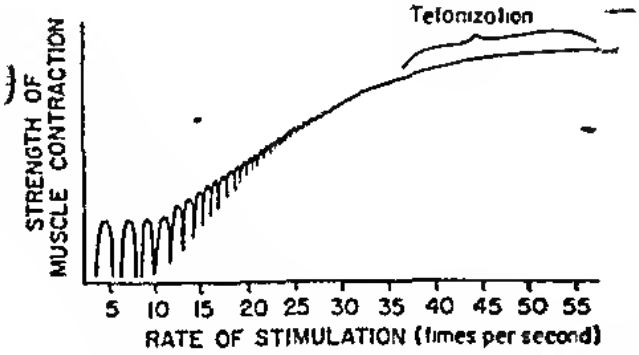


Figure 11–15. Wave summation and tetanization.

SKELETAL MUSCLE TONE

Even when muscles are at rest, a certain amount of tautness usually remains. This residual degree of contraction in skeletal muscle is called *muscle tone*. Since skeletal muscle fibers do not contract without an actual action potential to stimulate the fibers except in certain pathological conditions, it is believed that skeletal muscle tone results entirely from nerve impulses coming

rom the spinal cord. These in turn are controlled partly by impulses transmitted from the brain to the appropriate anterior motor neurons and partly by impulses that originate in *muscle spindles* located in the muscle itself.

Muscle spindles are sensory receptors that exist throughout essentially all skeletal muscles to detect the degree of muscle contraction. These will be discussed in detail in Chapter 51, but, briefly, they transmit impulses almost continually through the posterior roots into the spinal cord, where they excite the anterior motor neurons, which in turn provide nerve stimuli for muscle tone. Simply cutting the posterior roots, thereby blocking the muscle spindle impulses, usually reduces muscle tone to such a low level that the muscle becomes almost completely flaccid.

Many other neurogenic factors, originating especially in the brain, enter into the control of muscle tone These will be discussed in relation to muscle spindle and spinal cord function in Chapter 51.

MUSCLE FATIGUE

Prolonged and strong contraction of a muscle leads to the well-known state of muscle fatigue. Studies in athletes have shown that muscle fatigue increases in almost direct proportion to the rate of depletion of muscle glycogen. Therefore, most fatigue probably results simply from inability of the contractile and metabolic processes of the muscle fibers to continue supplying the same work output. However, experiments have also shown that transmission of the nerve signal through the neuromuscular junction can occasionally diminish following prolonged muscle activity, thus further diminishing muscle contraction.

Interruption of blood flow through a contracting muscle leads to almost complete muscle fatigue in a minute or more because of the obvious loss of nutrient supply—especially loss of oxygen.

THE LEVER SYSTEMS OF THE BODY

Muscles obviously operate by applying tension to their points of insertion into bones, and the bones in turn form various types of lever systems Figure 11–16 illustrates the lever system activated by the biceps muscle to lift the forearm If we assume that a large biceps muscle has a cross-sectional area of 6 square inches, then the maximum force of contraction would be about 300 pounds. When the forearm is exactly at right angles

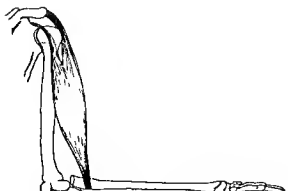


Figure 11–16. The lever system activated by the biceps muscle.

with the upper arm, the tendon attachment of the biceps is about 2 inches anterior to the fulcrum at the elbow, and the total length of the forearm lever is about 14 inches. Therefore, the amount of lifting power that the biceps would have at the hand would be only one seventh of the 300 pounds force, or about 43 pounds. When the arm is in the fully extended position the attachment of the biceps is much less than 2 inches anterior to the fulcrum, and the force with which the forearm can be brought forward is much less than 43 pounds.

In short, an analysis of the lever systems of the body depends on (a) a discrete knowledge of the point of muscle insertion and (b) its distance from the fulcrum of the lever, as well as (c) the length of the lever arm and (d) the position of the lever. Obviously, many different types of movement are required in the body, some of which need great strength and others large distances of movement. For this reason there are all varieties of muscles; some are long and contract a long distance and some are short but have large cross-sectional areas and therefore can provide extreme strengths of contraction over short distances The study of different types of muscles, lever systems, and their movements is called *kinesiology* and is a very important phase of human physioanatomy.

Accommodation of Muscle Length to the Length of the Lever System. If a bone is broken and then heals in a shortened state, the force of contraction of the muscles lying along this broken bone would obviously become decreased because of the shortened lengths of muscles. However, muscles shortened in this manner undergo *physical shortening* during the next few weeks This results from actual loss of sarcomeres at the ends of the muscle fibers. In this way the new muscle length becomes approximately equal to that required for optimal force of contraction. Conversely, if muscles are overstretched for weeks at a time, new sarcomeres develop at the ends of the fibers, again re-establishing optimal force of contraction

The same shortening process also occurs in muscles of limbs immobilized for several weeks in casts if the muscles during this time are in a shortened position. When the cast is removed, the muscles must often be restretched over a period of weeks before full mobility is restored

SPECIAL FEATURES AND ABNORMALITIES OF SKELETAL MUSCLE FUNCTION

MUSCLE HYPERTROPHY

Forceful muscular activity causes the muscle size to increase, a phenomenon called hypertrophy. Most of the hypertrophy results from increase in the diameters of the fibers already present in the muscle, but the numbers of fibers probably also can increase to a slight extent by splitting of fibers already present; this is called *hyperplasia*. As the diameters of the muscle fibers increase, the sarcoplasm increases, and the fibers gain in various nutrient and intermediary metabolic substances, such as adenosine triphosphate, phosphocreatine, glycogen, intracellular lipids, and even many additional mitochondria. It is likely that the myofibrils also increase in size and perhaps in numbers as well, but this has not been proved Briefly, muscular hyper-

trophy increases both the motive power of the muscle and the nutrient mechanisms for maintaining increased motive power.

Weak muscular activity, even when sustained over long periods of time, does not result in significant hypertrophy. Instead, hypertrophy results mainly from *very* forceful muscle activity, though the activity might occur for only a few minutes each day. For this reason, strength can be developed in muscles much more rapidly when "resistive" or "isometric" exercise is used rather than simply prolonged mild exercise. Indeed, essentially no enlargement of the muscle fibers occurs unless the muscle contracts to at least 75 per cent of its maximum tension.

On the other hand, prolonged muscle activity does increase muscle endurance, causing increases in the oxidative enzymes, myoglobin, and even blood capillaries—all of which are essential to increased muscle metabolism.

MUSCLE ATROPHY

Muscle atrophy is the reverse of muscle hypertrophy. It results any time a muscle is not used or even when a muscle is used only for very weak contractions. Atrophy is particularly likely to occur when limbs are placed in casts, thereby preventing muscular contraction. As little as one month of disuse can sometimes decrease the muscle size to one-half normal.

Atrophy Caused by Muscle Denervation. When a muscle is denervated it immediately begins to atrophy, and the muscle continues to decrease in size for several years. If the muscle becomes reinnervated during the first three to four months, full function of the muscle usually returns, but after four months of denervation some of the muscle fibers usually will have degenerated. Reinnervation after two years rarely results in return of any function at all. Pathological studies show that the muscle fibers have by that time been replaced by fat and fibrous tissue.

PREVENTION OF MUSCLE ATROPHY BY ELECTRICAL STIMULATION

Strong electrical stimulation of denervated muscles, particularly when the resulting contractions occur against loads, will delay and in some instances prevent muscle atrophy despite denervation. This procedure is used to keep muscles alive until reinnervation can take place.

PHYSICAL CONTRACTURE OF MUSCLE FOLLOWING DENERVATION

When a muscle is denervated, its fibers tend to shorten if the muscle is kept in a shortened position, and even the associated nerves and fasciae shorten. That is, unless continual movement keeps stretching the muscle and other structures, they will creep toward a shortened length. This is one of the most difficult problems in the treatment of patients with denervated muscles, such as occur in poliomyelitis or nerve trauma. Unless passive stretching is applied daily to the muscles, they may become so shortened that even when reinnervated they will be of little value. But, more important, the shortening can often result in extremely contorted positions of different parts of the body.

RIGOR MORTIS

Several hours after death all the muscles of the body go into a state of *contracture* called rigor mortis; that is, the muscle contracts and becomes rigid even without action potentials. This rigidity is caused by loss of all the ATP, which is required to cause separation of the cross-bridges from the actin filaments during the relaxation process. The muscles remain in rigor until the muscle proteins are destroyed, which usually results from autolysis caused by enzymes released from the lysosomes some 15 to 25 hours later.

FAMILIAL PERIODIC PARALYSIS

Occasionally, a hereditary disease called familial periodic paralysis occurs. In persons so afflicted, the extracellular fluid potassium concentration periodically falls to very low levels, causing various degrees of paralysis. The paralysis is caused in the following manner: The initial decrease in extracellular fluid potassium causes an increase in the gradient of potassium from the inside of the muscle fiber to the outside, and this in turn increases the muscle fiber membrane potential to a value greater than normal, an effect called *hyperpolarization*. For a long time it was believed that this hyperpolarization made the muscle fiber membranes less excitable than usual and caused the paralysis. However, with modern methods of study, it has been learned that this initial hyperpolarization is not the culprit. Instead when the extracellular fluid potassium concentration falls even lower (below about 3 mEq/liter), this low potassium level has the direct effect of decreasing the permeability of the potassium channels. Therefore, potassium now contributes much less to the membrane potential than normally, so that the intracellular membrane potential *decreases* below normal, often decreasing low enough that many or most of the sodium channels become inactivated. It will be recalled from the previous chapter that inactivation of the sodium channels prevents the transmission of action potentials; therefore this can lead to paralysis.

THE ELECTROMYOGRAM

Each time an action potential passes along a muscle fiber a small portion of the electrical current spreads away from the muscle as far as the skin. If many muscle fibers contract simultaneously, the summated electrical potentials at the skin may be very great. By placing two electrodes on the skin or inserting needle electrodes into the muscle, an electrical recording called the *electromyogram* can be made when the muscle is stimulated. Figure 11–17 illustrates a typical electromyographic recording from the gastrocnemius muscle during a moderate contraction. Electromyograms are frequently used clinically to discern abnormalities of muscle excitation.

Figure 11–17. Electromyogram recorded during contraction of the gastrocnemius muscle.

o such abnormalities are muscle *fasciculation* and
rillation.

Muscle Fasciculation. When an abnormal impulse
urs in a motor nerve fiber its whole motor unit
ntracts. This often causes sufficient contraction in the
uscle that one can see a slight ripple in the skin over
e muscle. This process is called fasciculation.
Fasciculation occurs especially following destruction
anterior motor neurons in poliomyelitis or following
umatic interruption of a nerve. As the peripheral
rve fibers die, spontaneous impulses are generated
ring the first few days, and fasciculatory muscle
vements result in the muscle. Typical electromy-
raphic records of weak periodic potentials can be
tained from the skin overlying the muscle.
Muscle Fibrillation. After all nerves to a muscle
ve been destroyed and the nerve fibers themselves
ve become nonfunctional, which requires three to five
, spontaneous impulses begin to appear in the de-
ervated muscle fibers. At first, these occur at a rate of
ce every few seconds, but, after a few more days or a
w weeks, the impulses become as rapid as three to ten
nes per second. Thus, skeletal muscle fibers, several
ys after loss of their innervation, develop an intrinsic
ythmicity. This is associated with, and probably
used by, the spread of large numbers of acetylcholine
ceptor proteins over the surfaces of the muscle fibers,
hich is believed to increase the permeability of the
er membranes. After several more weeks the muscle
bers atrophy to such an extent that the fibrillatory
npulses finally cease. To record an electromyogram of
orillation, minute bipolar needle electrodes must be
serted into the muscle belly itself because adjacent
uscle fibers do not fire simultaneously and, therefore,
) not summate. As a result, the potentials are not
rong enough to record from the surface of the skin.

EFERENCES

guayo A. J, and Karpati, G (eds) Current Topics in Nerve and Muscle Research New York, Elsevier/North-Holland, 1979
lexander, R. M Mechanics of skeleton and tendons *In* Brooks, V B (ed) Handbook of Physiology Sec. 1, Vol. II. Bethesda, American Physiological Society, 1981, p. 17
merican Physiological Society Skeletal Muscle Baltimore, Waverly Press, 1983
rmstrong, R. B Mechanisms of exercise-induced delayed onset muscular soreness. A brief review *Med. Sci Sports Exerc*, *16* 529, 1984.
arany, M, and Barany, K Phosphorylation of the myofibrillar proteins *Annu Rev Physiol*, *42* 275, 1980
uchthal, F., and Schmalbruch, H Motor unit of mammalian muscle *Physiol Rev*, *60* 90, 1980
urke, R. E Motor units. Anatomy, physiology, and functional organization *In* Brooks, V B (ed) Handbook of Physiology Sec. 1, Vol. II Bethesda, American Physiological Society, 1981, p. 345
aruso, G, (ed) Electromyography in the Diagnosis and Management of Peripheral Nerve Injuries Minneapolis, J K. Burgess, 1983
isenberg, E., and Greene, L. E. The relation of muscle biochemistry to muscle physiology *Annu Rev Physiol*, *42* 293, 1980
'sbiato, A, and Fabiato F Calcium and cardiac excitation-contraction coupling *Annu Rev Physiol*, *41* 473, 1979
reund, H. J., Motor unit and muscle activity in voluntary motor control *Physiol Rev*, *63* 387, 1983
alley, F. M: Human Movement. An Introductory Text for Physiotherapy Students New York, Churchill Livingstone, 1982.
lassalbach, W, and Oetliker H· Energetics and electrogenesity of the sarcoplasmic reticulum calcium pump *Annu Rev Physiol*, *45*.325, 1983
Hockey, R Stress and Fatigue in Human Performance. New York, John Wiley & Sons, 1983
Homsher, E, and Kean, C J . Skeletal muscle energetics and metabolism. *Annu Rev.Physiol*, *40* 93, 1978
Huxley, A F, and Gordon, A M : Striation patterns in active and passive shortening of muscle *Nature (Lond.)*, *193* 280, 1962
Huxley, H. E. Muscular contraction and cell motility *Nature*, *243* 445, 1973
Huxley, H. E, and Faruqi, A. R · Time-resolved x ray diffraction studies on vertebrate striated muscle *Annu Rev Biophys Bioeng.*, *12*.381, 1983.
Ikemoto, N Structure and function of the calcium pump protein of sarcoplasmic reticulum *Annu Rev Physiol*, *44* 297, 1982
Ingels, N B, Jr (ed) The Molecular Basis of Force Development in Muscle Palo Alto, Cal, Palo Alto Medical Research Foundation, 1979.
Johnson, E. W. (ed) Practical Electromyography. Baltimore, Williams & Wilkins, 1979.
Jolesz, F, and Sreter, F A Development, innervation, and activity-pattern induced changes in skeletal muscle. *Annu. Rev Physiol*, *43* 531, 1981
Knochel, J P Rhabdomyolysis and myoglobinuria *Annu Rev Med.*, *33* 435, 1982
Korn, E D Actin polymerization and its regulation by proteins from nonmuscle cells *Physiol Rev*, *62* 672, 1982
Lamb, D H. Physiology of Exercise Responses and Adaptations. New York, Macmillan Publishing Co, 1984
Lymn, R W Kinetic analysis of myosin and actomyosin ATPase. *Annu Rev Biophys Bioeng*, *6* 145, 1979
Martonosi, A The development of sarcoplasmic reticulum membranes *Annu Rev Physiol*, *44* 377, 1982
Martonosi, A N Mechanisms of Ca^{++} release from sarcoplasmic reticulum of skeletal muscle *Physiol Rev*, *64* 1240, 1984
McLachlan, A D Structural implications of the myosin amino acid sequence *Annu Rev Biophys Bioeng*, *13* 167, 1984
Morgan, D L, and Proske, U Vertebrate slow muscle Its structure, pattern of innervation, and mechanical properties. *Physiol Rev*, *64* 103, 1984
Mortimore, J T Motor prostheses *In* Brooks, V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 155
Northrip, J W, *et al* Introduction to Biomechanic Analysis of Sport. Dubuque, Iowa. W C Brown Co, 1979
Partridge, L. D, and Benton, L A. Muscle, the motor *In* Brooks, V B. (ed) Handbook of Physiology Sec 1, Vol II. Bethesda, American Physiological Society, 1981, p 43
Pearson, M L (ed) Muscle Development Molecular and Cellular Control. Cold Spring Harbor, N Y , Cold Spring Harbor, 1982
Pollack, G H The cross-bridge theory *Physiol Rev*, *63* 1049, 1983
Porter, R (ed) Human Muscle Fatigue. Woodstock, N.Y., Beekman Publishers, 1981
Rash, J Neuromuscular Atlas New York, Prager Publishers, 1984
Regart, R. Sodium channels in nerve and muscle membrane *Annu Rev Physiol*, *43* 711, 1981
Schneider, M F Membrane charge movement and depolarization contraction coupling *Annu Rev Physiol*, *43* 507, 1981
Schubert, D Developmental Biology of Cultured Nerve, Muscle and Glia New York, John Wiley & Sons, 1984
Spolin, R A Ion Transport in Skeletal Muscle New York, John Wiley & Sons 1982
Stefani, E, and Chiarandini, D J · Ionic channels in skeletal muscle *Annu. Rev Physiol*, *44* 357, 1982
Sugi, H, and Pollack, G H. (eds.) Cross-Bridge Mechanism in Muscle Contraction Baltimore, University Park Press, 1979.
Taylor, C R, and Heglund, N. C.: Energetics and mechanics of terrestrial locomotion *Annu. Rev Physiol.*, *44* 97, 1982
Tregear, R T, and Marston, S. B.: The crossbridge theory. *Annu. Rev. Physiol*, *41* 723, 1979.
Walton, J. A: Disorders of Voluntary Muscle. New York, Churchill Livingstone, 1981.
Wiesendanger, M, and Miles, T. S · Ascending pathway of low-threshold muscle afferents to the cerebral cortex and its possible role in motor control *Physiol Rev*, *62* 1234, 1982.
Winegrad, S: Regulation of cardiac contractile proteins Correlations between physiology and biochemistry. *Circ Res*, *55* 565, 1984
Wray, J. S, and Holmes, K. C.: X-ray diffraction studies of muscle *Annu. Rev Physiol*, *43* 553, 1981
Zak, R., and Rabinowitz, M · Molecular aspects of cardiac hypertrophy *Annu. Rev Physiol*, *41* 539, 1979

12

Neuromuscular Transmission; Function of Smooth Muscle

TRANSMISSION OF IMPULSES FROM NERVES TO SKELETAL MUSCLE FIBERS: THE NEUROMUSCULAR JUNCTION

The skeletal muscles are innervated by large, myelinated nerve fibers that originate in the large motoneurons of the anterior horns of the spinal cord. It was pointed out in the previous chapter that each nerve fiber normally branches many times and stimulates from three to several hundred skeletal muscle fibers. The nerve ending makes a junction, called the *neuromuscular junction*, with the muscle fiber approximately at the fiber's midpoint so that the action potential in the fiber travels in both directions. With the excep of about 2 per cent of the muscle fibers ther only one such junction per muscle fiber.

Physiologic Anatomy of the Neuromuscu Junction. Figure 12–1, Parts A and B, illustr the neuromuscular junction between a large, m linated nerve fiber and a skeletal muscle fi The nerve fiber branches at its end to forr complex of branching nerve *terminals* called *end-plate*, which invaginates into the muscle fi but lies entirely outside the muscle fiber plas membrane. The entire structure is covered by or more Schwann cells that insulate the end-pl from the surrounding fluids.

Figure 12–1C shows an electron micrograp sketch of the junction between a single-bra

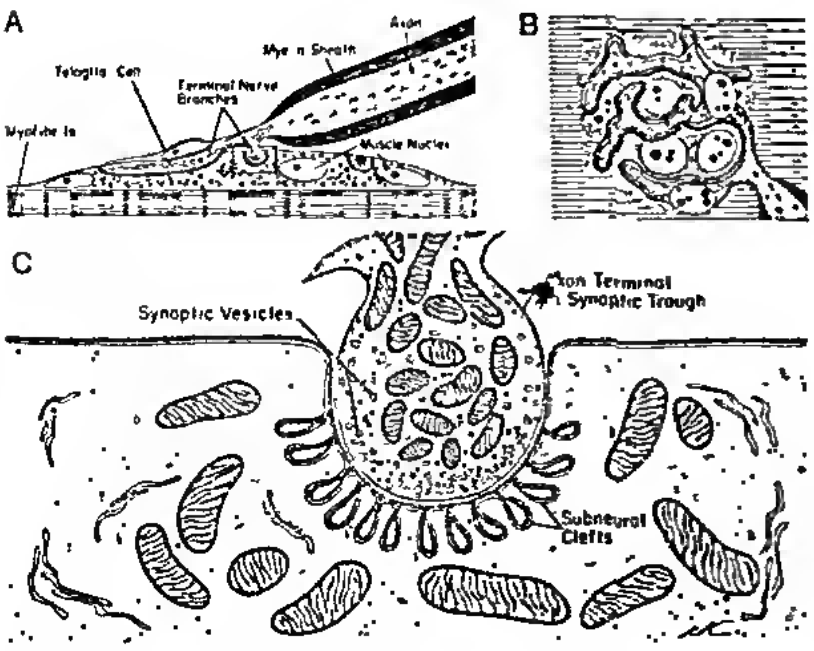


Figure 12–1. Different views of the motor plate A. Longitudinal section through the plate B. Surface view of the end plate C. Elec micrographic appearance of the contact point tween one of the axon terminals and the mu fiber membrane, representing the rectangular shown in A. (From Bloom and Fawcett, as mod from R. Couteaux A Textbook of Histology adelphia, W B Saunders Company, 1975)

xon terminal and the muscle fiber membrane. he invagination of the membrane is called the *naptic gutter* or *synaptic trough*) and the space etween the terminal and the fiber membrane is alled the *synaptic cleft*. The synaptic cleft is 20 to 0 nanometers wide and is occupied by a basal amina, which is a thin layer of spongy reticular bers through which diffuses extracellular fluid. t the bottom of the gutter are numerous smaller olds of the muscle membrane called *subneural lefts*, which greatly increase the surface area at hich the synaptic transmitter can act. In the xon terminal are many mitochondria that supply nergy mainly for synthesis of the excitatory transatter *acetylcholine* that, in turn, excites the muscle fiber. The acetylcholine is synthesized in the ytoplasm of the terminal but is rapidly absorbed nto many small synaptic vesicles, approximately 00,000 of which are normally in the terminals of single end-plate. Attached to the matrix of the asal lamina are large quantities of the enzyme *cetylcholinesterase*, which is capable of destroying cetylcholine, to be explained in further detail.

Secretion of Acetylcholine by the Axon Terminals. When a nerve impulse reaches the neuromuscular junction, about 300 vesicles of acetylcholine are released by the terminals into the ynaptic clefts between the terminals and the muscle fiber membrane. This results from movement of calcium ions from the extracellular fluid into he membranes of the terminals when the action otential depolarizes their membranes. The calcium ions cause the vesicles of acetylcholine to rupture through the membrane. In the absence of calcium or in the presence of excess magnesium, he release of acetylcholine is greatly depressed.

Destruction of the Released Acetylcholine by Cholinesterase. Within approximately 1 millisecond after acetylcholine is released by the axon erminal, much of it has already diffused out of he synaptic gutter and no longer acts on the muscle fiber membrane, and virtually all the remainder is destroyed by the acetylcholinesterase n the basal lamina lying between the nerve terminal and the subneural clefts. The very short period of time that the acetylcholine remains in contact with the muscle fiber membrane—about 1 millisecond—is almost always sufficient to excite the muscle fiber, and yet the rapid removal of the acetylcholine prevents re-excitation after the muscle fiber has recovered from the first action potential.

Effect of Acetylcholine in Opening Acetylcholine-Gated Ion Channels. Even though the acetylcholine released into the cleft between the end-plate and the muscle membrane lasts for only a very small fraction of a second, during this time the permeability of the muscle membrane to positive ions in the cleft increases several thousand-fold. This is because of the opening of *acetylcholine-gated ion channels*. These channels are composed basically of a large, transmembrane, channel protein molecule with a molecular weight of about 240,000. On its external surface is a gate similar to that of the sodium channels discussed in Chapter 10. However, it is a *ligand-activated gate* rather than a voltage-activated gate, responding to acetylcholine as the ligand. The surface-protruding portion of this channel protein acts as a receptor for acetylcholine, and acetylcholine in turn causes a conformational change in the channel molecule to open the gate for about 1 millisecond.

The acetylcholine channels have a diameter of about 0.65 nanometers, which is large enough to allow all the important positive ions—Na^+, K^+, and Ca^{++}—to move easily through the channels. (Negative ions such as chloride ions do not pass through this channel because of strong negative charges in the wall of the channel.)

Yet, from a practical point of view only sodium ions flow through the channels, for the following reasons: First, the normal resting potential of the muscle fiber membrane is almost equal to the Nernst potential for potassium ions; therefore, opening the channels will not allow significant numbers of potassium ions to flow through the membrane—that is, the potassium concentration gradient is already approximately in equilibrium with the voltage gradient across the channels. On the other hand, both sodium and calcium ions have far greater concentrations outside the nerve fiber than inside, and their Nernst potentials are of opposite polarity to the resting membrane potential. Therefore, opening the acetylcholine channels allows both these ions to flow rapidly to the interior of the muscle fiber. However, the concentration of calcium ions in the extracellular fluid is less than one fiftieth the concentration of sodium ions, and their mobility is also less. Therefore, in essence, opening the acetylcholine channels allows very rapid influx of sodium ions to the interior of the fiber and little else. Nevertheless, even 1 millisecond influx of the sodium ions is still enough to excite an action potential in the muscle fiber.

The "End-Plate Potential" and Excitation of the Skeletal Muscle Fiber. The sudden insurgence of sodium ions into the muscle fiber when the acetylcholine channels open causes the membrane potential in the *local area of the end-plate* to increase in the positive direction as much as 50 to 75 millivolts, creating a *local potential* called the *end-plate potential*. If we recall from the previous chapter that a sudden increase in membrane potential of more than 15 to 30 millivolts is sufficient to initiate the positive feedback effect of sodium channel activation, one can understand that the end-plate potential created by the acetylcholine stimulation is normally far greater than enough to initiate an action potential in the muscle fiber.

Figure 12–2 illustrates the principle of an end-plate potential initiating the action potential. In this figure are shown three separate end-plate potentials. End-plate potentials A and C are too weak to elicit an action potential, but they do nevertheless give the weak local potentials recorded in the figure. In contrast, end-plate

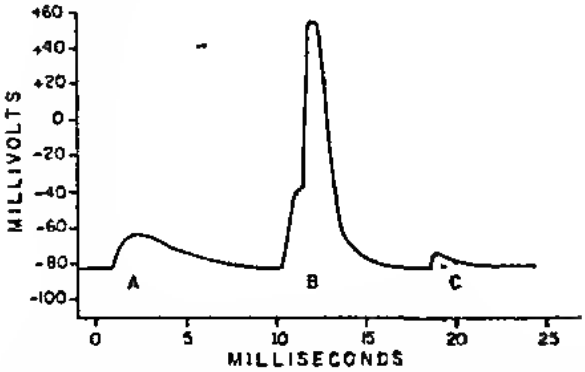


Figure 12–2. End-plate potentials. *A*, A weakened end-plate potential recorded in a curarized muscle, too weak to elicit an action potential. *B*, normal end-plate potential eliciting a muscle action potential, and *C*, weakened end-plate potential caused by botulinum toxin that decreases end-plate release of acetylcholine, again too weak to elicit a muscle action potential

potential B is much stronger and causes enough sodium channels to activate so that the self-regenerative effect of more and more sodium ions flowing to the interior of the fiber initiated an action potential. The weak end-plate potential at point A was caused by poisoning the muscle fiber with *curare*, a drug that blocks the gating action of acetylcholine on the acetylcholine channels, probably by competing with the acetylcholine for the acetylcholine receptor site. The weak end-plate potential at point C resulted from the effect of botulinum toxin, a bacterial toxin that decreases the release of acetylcholine by the nerve terminals.

"Safety Factor" for Transmission at the Neuromuscular Junction; Fatigue of the Junction. Ordinarily, each impulse that arrives at the neuromuscular junction creates an end-plate current flow about three to four times that required to stimulate the muscle fiber. Therefore, the normal neuromuscular junction is said to have a very high *safety factor*. However, artificial stimulation of the nerve fiber at rates greater than 100 times per second for several minutes often diminishes the number of vesicles of acetylcholine released with each impulse so much that impulses then fail to pass into the muscle fiber. This is called *fatigue* of the neuromuscular junction, and it is analogous to fatigue of the synapse in the central nervous system. Under normal functioning conditions, fatigue of the neuromuscular junction probably occurs only at the most exhausting levels of muscular activity.

Molecular Biology of Acetylcholine Formation and Release

The neuromuscular junction is a type of *synapse*, which is defined as a junction between two excitable cells that allows transmission of a signal from the first cell to the next. We shall see in Chapter 46 that literally trillions of synapses occur in the central nervous system between one neuron and the next, these function in much the same way as the neuromuscular junction. However, the central nervous system synapses are so small that it has not been possible to study the precise events of signal transmission in these. On the other hand, the neuromuscular junctions are large enough to be easily studied, and we know much about the formation and release of acetylcholine at this junction. This occurs in the following stages.

(1) Very small vesicles, about 40 nanometers in [illegible] are formed by the Golgi apparatus in the cell body [illegible] the motor neuron in the spinal cord. These vesicles [illegible] then transported by "streaming" of the axopl[illegible] through the core of the axon from the central cell bod[illegible] to the neuromuscular junction at the tips of the ner[illegible] fibers. About 300,000 of these small vesicles collect i[illegible] the nerve terminals of a single end-plate.

(2) Acetylcholine is synthesized in the cytosol of t[illegible] terminal nerve fibers but is then transported throug[illegible] the membranes of the vesicles to their interior, whe[illegible] it is stored in highly concentrated form, with abo[illegible] 10,000 molecules of acetylcholine in each vesicle.

(3) Under resting conditions an occasional vesic[illegible] fuses with the surface membrane of the nerve termin[illegible] and releases its acetylcholine into the synaptic gute[illegible] When this occurs, a so-called *miniature end-plate* pote[illegible] tial, about 1 millivolt in intensity and lasting for a fe[illegible] milliseconds, occurs in the local area of the muscle fib[illegible] because of the action of this "packet" of acetylcholine.

(4) When an action potential arrives at the ner[illegible] terminal, this opens many calcium channels in th[illegible] membrane of the terminal because this terminal has a[illegible] abundance of voltage-gated calcium channels. As a r[illegible] ault, the calcium ion concentration in the termin[illegible] increases about a hundred-fold, which in turn increas[illegible] the rate of fusion of the acetylcholine vesicles with [illegible] terminal membrane by about 10,000-fold. As each ve[illegible] cle fuses, its outer surface ruptures through [illegible] membrane, thus causing exocytosis of acetylcholine [illegible] the synaptic cleft. Usually about 200 to 300 vesic[illegible] rupture with each action potential. This ent[illegible] of events occurs in approximately 5 milliseconds.

(5) After each vesicle has released its acetylcholin[illegible] the membrane of the vesicle becomes part of the [illegible] membrane. However, the number of vesicles availabl[illegible] in the nerve ending is sufficient only to allow transmi[illegible] sion of about 1000 nerve impulses. Therefore, for conti[illegible] ued function of the neuromuscular junction, the vesicle[illegible] need to be retrieved from the nerve membrane. Retrie[illegible] is achieved by the process of *pinocytosis* that was e[illegible] plained in Chapter 2. Within a few seconds after th[illegible] action potential is over, "coated pits" appear on th[illegible] surface of the terminal nerve membrane, caused b[illegible] contractile proteins of the cytosol massing at interm[illegible] tent points underneath the membrane. The proteins [illegible] the original vesicles are attracted to these pits, an[illegible] within about 20 seconds the contractile proteins contrac[illegible] and cause the pits to break away to the interior of th[illegible] membrane, thus forming new vesicles. Within anothe[illegible] few seconds, acetylcholine is transported to the interi[illegible] of these vesicles, and they are then ready for a ne[illegible] cycle of acetylcholine release.

Drugs That Affect Transmission at the Neuromuscular Junction

Drugs That Stimulate the Muscle Fiber by Acetylcholine-Like Action. Many different compounds, including *methacholine*, *carbachol*, and *nicotine*, have the same effect on the muscle fiber as does acetylcholine. The difference between these drugs and acetylcholine is that they are not destroyed by cholinesterase or are destroyed very slowly, so that when once applied to the muscle fiber the action persists for many minutes to several hours. Moderate quantities of the above three drugs applied to a muscle fiber cause localized areas

olarization, and every time the muscle fiber becomes larized elsewhere, these depolarized areas, by virtue their leaking ions, cause new action potentials, eby causing a state of spasm. On the other hand, n extreme doses of these drugs are used two separate cts occur that now cause a state of flaccid paralysis ner than spasm. These effects are: (1) Prolonged on of acetylcholine or acetylcholine-like drugs causes acetylcholine-gated ion channels to become desenzed to these drugs, which is analogous to the inactiion process of sodium channels as discussed in the vious chapter. (2) The moderate degree of depolarion that still persists also causes inactivation of the ium channels so that they now cannot participate in initiation of action potentials.

Drugs That Block Transmission at the Neuroscular Junction. A group of drugs, known as the *curiform drugs*, can prevent passage of impulses from end-plate into the muscle. Thus, D-tubocurarine cts the membrane, probably by competing with acecholine for the receptor sites of the membrane, so t the acetylcholine cannot increase the permeability the acetylcholine channels sufficiently to initiate a olarization wave.

Drugs That Stimulate tha Neuromuscular Junc-n by Inactivating Acetylcholinesterase. Three ticularly well-known drugs, *neostigmine, physostigne,* and *diisopropyl fluorophosphate,* inactivate acecholinesterase so that the cholinesterase normally in synapses will not hydrolyze the acetylcholine resed at the end-plate. As a result, acetylcholine inases in quantity with successive nerve impulses so t extrsme amounts of acetylcholine can accumulate l then repetitively stimulate the muscle fiber. This ises *muscular spasm* when even a few nerve impulses ch the muscle; this can cause death due to laryngeal asm, which smothers the person.

Neostigmine and physostigmine combine with acetylolinesterase to inactivate it for several hours, after ich they are displaced from the acetylcholinesterase that it once again becomes active. On the other hand, sopropyl fluorophosphate, which has military potenl as a very powerful "nerve" gas, actually inactivates etylcholinesterase for several weeks, which makes this particularly lethal drug.

YASTHENIA GRAVIS

The disease *myasthenia gravis*, which occurs in about e of every 20,000 persons, causes the person to become ralyzed because of inability of the neuromuscular nctions to transmit signals from the nerve fibers to e muscle fibers Pathologically, the number of subural clefts in the synaptic gutter is reduced and the naptic cleft itself is widened as much as 50 per cent. lso, antibodies that attack the acetylcholine-gated ansport proteins have been demonstrated in the bloods these patients. Therefore, it is believed that myasenia gravis is an autoimmune disease in which paents have developed antibodies against their own acecholine-activated ion channels

Regardless of the cause, the end-plate potentials deloped in the muscle fibers are too weak to stimulate e muscle fibers adequately. If the disease is intense ough, the patient dies of paralysis—in particular, of ralysis of the respiratory muscles. However, the disse can usually be ameliorated with several different ugs, as follows:

Treatment with Drugs. When a patient with myasthenia gravis is treated with a drug, such as neostigmine, that is capable of inactivating acetylcholinesterase, the acetylcholine secreted by the end-plate is not destroyed immediately. If a sequence of nerve impulses arrives at the end-plate, the quantity of acetylcholine present at the membrane increases progressively until finally the end-plate potential caused by the acetylcholine rises above threshold value for stimulating the muscle fiber. Thus, it is sometimes possible by diminishing the quantity of acetylcholinesterase in the muscles of a patient with myasthenia gravis to allow even the inadequate quantities of acetylcholine secreted at the end-plates to effect almost normal muscular activity.

CONTRACTION OF SMOOTH MUSCLE

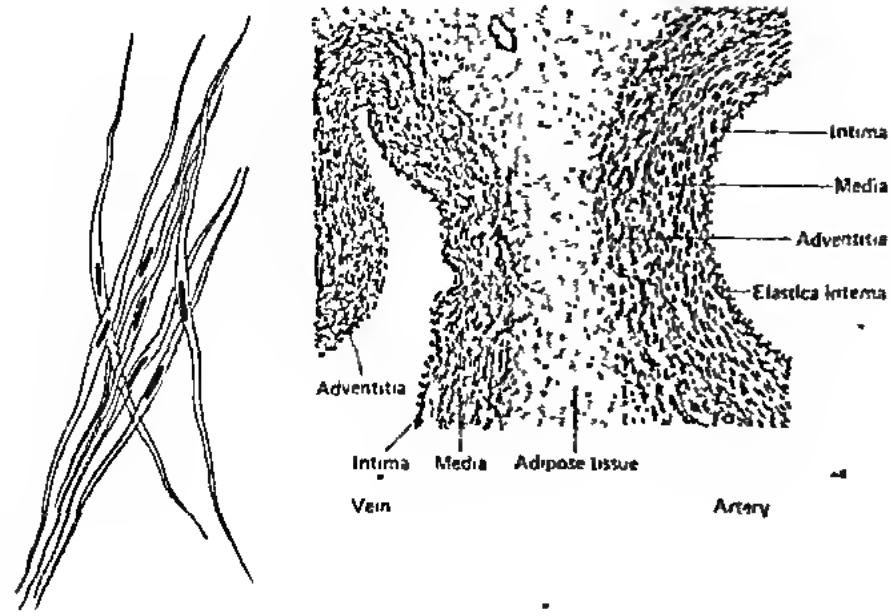
In the previous chapter and thus far in the present chapter, the discussion has been concerned with skeletal muscle. We now turn to smooth muscle, which is composed of far smaller fibers—usually 2 to 5 microns in diamater and only 50 to 200 microns in length—in contrast to the skeletal muscle fibers that ara as much as 20 timea aa large (in diameter) and thousanda of times aa long. Nevertheless, many of the principles of contraction apply to smooth muscle the aame as to skeletal muscle. Most importantly, essentially the same chemical abstances causa contraction in smooth muscle as in skeletal muscle, but the physical arrangement of smooth muscle fibers is entirely different, as we shall see

TYPES OF SMOOTH MUSCLE

The smooth muscle of each organ is distinctive from that of most other organs in several different ways: physical dimensions, organization into bundles or sheets, response to different types of stimuli, characteristics of innervation, and function. Yet, for the sake of simplicity, smooth muscle can generally be divided into two major types, which are illustrated in Figure 12–3; *multiunit smooth muscle* and *visceral smooth muscle.*

Multiunit Smooth Muscle. This type of smooth muscle is composed of discrete smooth muscle fibers. Each fiber operates entirely independently of the others and is often innervated by a single nerve ending, as occurs for skeletal muscle fibers. Furthermore, the outer surfaces of these fibers, like those of skeletal muscle fibers, are covered by a thin layer of "basement membrane–like" substance, a glycoprotein that helps insulate the separate fibers from each other.

The most important characteristic of multiunit smooth muscle fibers is that their control is exerted mainly by nerve signals (even though they themselves usually do not generate action potentials, as we shall discuss later). This is in contrast to a major share of the control of visceral smooth muscle by non-nervous stimuli. An additional charac-



Multiunit smooth muscle fibers
A

Visceral smooth muscle
B

Figure 12–3. Visceral and multiunit smooth muscle fibers.

teristic is that they rarely exhibit spontaneous contractions.

Some examples of multiunit smooth muscle found in the body are the smooth muscle fibers of the ciliary muscle of the eye, the iris of the eye, the nictitating membrane that covers the eyes in some lower animals, the piloerector muscles that cause erection of the hairs when stimulated by the sympathetic nervous system, and the smooth muscle of many of the larger blood vessels.

Visceral Smooth Muscle. Visceral smooth muscle fibers are similar to multiunit fibers except that they are usually arranged in sheets or bundles and the cell membranes contact each other at multiple points to form many *gap junctions* through which ions can flow freely from one cell to the next. Thus the fibers form a *functional syncytium* that usually contracts in large areas at once. For this reason, this type of smooth muscle is also known as *single-unit* or *unitary smooth muscle*. This type of muscle is found in most of the organs of the body, especially in the walls of the gut, the bile ducts, the ureters, the uterus, and so forth.

When one portion of a visceral muscle tissue is stimulated, the action potential is conducted to the surrounding fibers by *direct electrical conduction*. That is, the action potential generated in one area of the muscle electrically excites the adjacent fibers without secretion of any transmitter substance. Instead, electrical current flows through the gap junctions so easily that action potentials spread directly from one smooth muscle fiber to the next almost as though cell membranes did not exist between the fibers.

THE CONTRACTILE PROCESS IN SMOOTH MUSCLE

The Chemical Basis for Smooth Muscle Contraction. Smooth muscle contains both *actin* and *myosin filaments*, having chemical characteristics similar to but not exactly the same as those of the actin and myosin filaments in skeletal muscle. Smooth muscle also contains *tropomyosin*, but it is doubtful whether troponin or a troponin-like substance exists in smooth muscle. This raises a question about the mechanism for control of smooth muscle contraction, which will be discussed in more detail in a subsequent section of this chapter.

Chemical studies have shown that actin and myosin derived from smooth muscle interact with each other in the same way that this occurs for actin and myosin derived from skeletal muscle. Furthermore, the contractile process is activated by calcium ions, and ATP is degraded to ADP to provide the energy for contraction.

On the other hand, there are major differences between the physical organization of smooth muscle and that of skeletal muscle, as well as differences in other aspects of smooth muscle function such as excitation-contraction coupling, control of the contractile process by calcium ions, duration of contraction, and amount of energy required for the contractile process.

The Physical Basis for Smooth Muscle Contraction. Smooth muscle does not have the same striated arrangement of the actin and myosin filaments as that found in skeletal muscle. And, for a long time, it was impossible to discern even in

electron micrographs any specific organization in the smooth muscle cell that could account for contraction. However, recent special electron micrographic techniques suggest the physical organization illustrated in Figure 12–4. This shows large numbers of actin filaments attached to so-called *dense bodies*. Some of these bodies in turn are attached to the cell membrane whereas others are located throughout the cell but are held in place by a scaffold of structural protein cross-attachments from one dense body to another. Interspersed among the actin filaments are a few thick filaments about 2.5 times the diameter of the thin actin filaments.[7] These are assumed to be myosin filaments. However, there are only one twelfth to one fifteenth as many of these "myosin filaments" as actin filaments.

Despite the relative paucity of myosin filaments, it is assumed that they have sufficient cross-bridges to attract the many actin filaments and cause contraction by the sliding filament mechanism in essentially the same way that this occurs in skeletal muscle. And it is especially interesting to note that the maximum strength of contraction of smooth muscle is approximately equal to that of skeletal muscle, about 3 kg per square centimeter of cross-sectional area of the muscle.

Slowness of Contraction and Relaxation of Smooth Muscle. Though each smooth muscle tissue in the body has its own characteristics quite distinctive from the others, a typical smooth muscle tissue will begin to contract 50 to 100 milliseconds after it is excited, and will reach full contraction about half a second later. Then the contraction declines in another 1 to 2 seconds, giving a total contraction time of 1 to 3 seconds, which is about 30 times as long as the single-twitch contraction of skeletal muscle. However, smooth muscle contractions as short as 0.2 second and as long as 30 seconds also occur.

A major share of the prolonged contractile state of smooth muscle seems to be caused by slowness of the chemical reactions that cause the contraction. For instance, assuming that smooth muscle contraction occurs by the same "walk-along" mechanism as that proposed for skeletal muscle, it has been calculated that the frequency of power strokes by the heads of the cross-bridges is only one tenth to one hundredth as rapid as in skeletal muscle. This is believed to result from the fact that the cross-bridge heads have far less ATPase activity than that exhibited by the cross-bridge heads in skeletal muscle.

Energy Required to Sustain Smooth Muscle Contraction. Measurements have shown that only one twentieth to one four-hundredth as much energy is required to sustain the same tension of contraction in smooth muscle as in skeletal muscle. This also is believed to result mainly from the very slow activity of the myosin ATPase, but also from the fact that there are far fewer myosin filaments in smooth muscle than in skeletal muscle.

This economy of energy utilization by smooth muscle is exceedingly important to overall function of the body, because organs such as the intestines, the urinary bladder, the gallbladder, and other viscera must maintain moderate degrees of muscle contractile tone day in and day out.

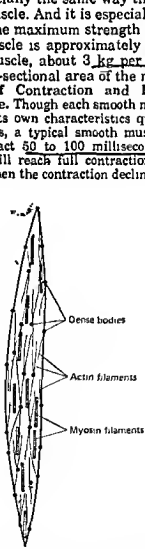


Figure 12–4. Arrangement of actin and myosin filaments in the smooth muscle cell. Note the attachment of the actin filaments to "dense bodies," some of which are themselves attached to the cell membrane.

MEMBRANE POTENTIALS AND ACTION POTENTIALS IN SMOOTH MUSCLE

Smooth muscle exhibits membrane potentials and action potentials similar to those that occur in skeletal muscle fibers. Furthermore, smooth muscle contraction can be elicited by depolarization of the membrane in the same way that contraction is initiated by depolarization of skeletal muscle fibers. However, there are both quantitative and qualitative differences in the membrane potentials and action potentials of smooth muscle that require special attention.

Membrane Potentials in Smooth Muscle. The quantitative value of the membrane potential of smooth muscle is variable from one type of smooth muscle to another, and it also depends on the momentary condition of the muscle. However, in the normal resting state, the membrane potential is usually about −50 to −60 millivolts, or about 30 millivolts less negative than in skeletal muscle.

Action Potentials in Visceral Smooth Muscle. Action potentials occur in visceral smooth muscle in the same way that they occur in skeletal muscle. However, action potentials probably do not normally occur in multiunit types of smooth muscle, as will be discussed in a subsequent section.

The action potentials of visceral smooth muscle occur in two different forms: (1) spike potentials and (2) action potentials with plateaus.

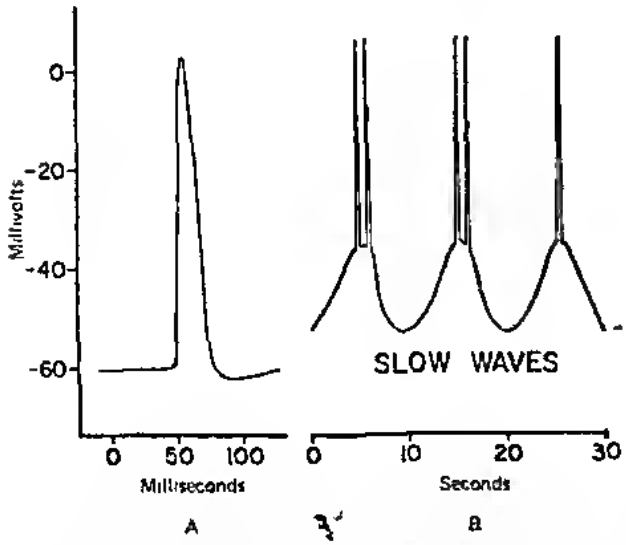


Figure 12–5. *A*, A typical smooth muscle action potential (spike potential) elicited by an external stimulus. *B*, A series of spike action potentials elicited by rhythmic, slow electrical waves occurring spontaneously in the smooth muscle wall of the intestine.

Spike Potentials. Typical spike action potentials, such as those seen in skeletal muscle, occur in most types of visceral smooth muscle. The duration of this type of action potential is 10 to 50 milliseconds, as illustrated in Figure 12–5A. Such action potentials can be elicited in many ways, such as by electrical stimulation, by the action of hormones on the smooth muscle, by the action of transmitter substances from nerve fibers, or as a result of spontaneous generation in the muscle fiber itself, as discussed below.

Action Potentials with Plateaus. Figure 12–6 illustrates an action potential with a plateau. The onset of this action potential is similar to that of the typical spike potential. However, instead of rapid repolarization of the muscle fiber membrane, the repolarization is delayed for several hundred to several thousand milliseconds. The importance of the plateau is that it can account for the prolonged periods of contraction that occur in at least some types of smooth muscle. This type of action potential often occurs in the ureter, in the uterus under some conditions, and in some types of vascular smooth muscle. (Also, this is the type of action potential seen in cardiac muscle fibers th have a prolonged period of contraction, as we sha discuss in the next two chapters.)

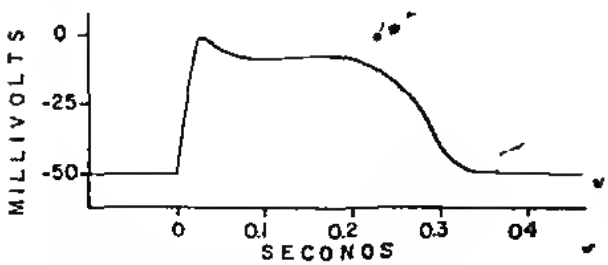


Figure 12–6. Monophasic action potential from a smooth muscle fiber of the rat uterus.

Importance of Calcium Channels in Gene ating the Smooth Muscle Action Potential. Th smooth muscle cell membrane has far more vol age-gated calcium channels than does skelet muscle, but far fewer sodium channels. Therefor sodium participates to a far less extent in th generation of the action potential in smooth musc than in skeletal muscle. Instead, flow of ca cium ions to the interior of the fiber is main responsible for the action potential. This occurs i the same self-regenerative way as occurs for th sodium channels in nerve fibers and in skelet muscle fibers. However, calcium channels ope many times more slowly than do sodium channel This accounts in large measure for the slow actio potentials of smooth muscle fibers.

Another important feature of calcium entry in the cells during the action potential is that th calcium acts directly on the smooth muscle act and myosin contractile mechanism to cause co traction, as we shall discuss later. Thus, the ca cium performs two tasks at once.

Slow Wave Potentials in Visceral Smoot Muscle and Spontaneous Generation of Actio Potentials. Some smooth muscle is self-excitator That is, action potentials arise within the smoot muscle itself without an extrinsic stimulus. Th is usually associated with a basic *slow wa rhythm* of the membrane potential. A typical slo wave of this type is illustrated in Figure 12–5 The slow wave itself is not an action potential. is not a self-regenerative process that sprea progressively over the membranes of the musc fibers. Instead, it is a local property of the smoot muscle fibers that make up the muscle mass.

The cause of the slow wave rhythm is as y unknown; one suggestion is that the slow wav are caused by waxing and waning of the pumpi of sodium outward through the muscle fiber me brane, the membrane potential becomes more ne ative when sodium is pumped rapidly and le negative when the sodium pump becomes le active. Another suggestion is that the condu tances of the ion channels increase and decrea rhythmically.

The importance of the slow waves lies in th fact that they can initiate action potentials. Th slow waves themselves cannot cause muscle co traction, but when the potential of the slow wav rises above the level of approximately −35 mill volts (the approximate threshold for eliciting ac tion potentials in most visceral smooth muscle an action potential develops and spreads over th visceral smooth muscle mass, and then contractio does occur. Figure 12–5B illustrates this effect showing that at each peak of the slow wave, on or more action potentials occur. This effect ca obviously promote a series of rhythmical contrac

of the smooth muscle mass. Therefore, the ow waves are frequently called *pacemaker waves.* his type of activity is especially prominent in bular types of smooth muscle masses, such as in e gut, the ureter, and so forth. In Chapter 63 we all see that this type of activity controls the ythmical contractions of the gut.

Excitation of Visceral Smooth Muscle by Stretch. When visceral smooth muscle is stretched ifficiently, spontaneous action potentials are usuly generated. These result from a combination of e normal slow wavo potentials plus a decrease the membrane potential caused by the stretch self. This response to stretch is an especially nportant function of visceral smooth muscle beause it allows a hollow organ that is excessively retched to contract automatically and therefore resist the stretch. For instonce, when the gut is verstretched by intestinal contents, a local autoatic contraction often sets up a peristaltic wave iat moves the contents away from the superretched intestino.

Depolarization of Multiunit Smooth Muscle Vithout Action Potentials. The smooth muscle bers of multiunit smooth muscle normally conact mainly in responso to nerve stimuli. The erve endiñga secrete acetylcholine in the case of ome multiunit amooth muscles and norepinephne in tho case of others. In both instances, these ransmitter substonces cause depolarization of the mooth muscle membrane, and this response in urn elicits tho contraction. However, action potenala most often do not develop. The reason for this that the fibers are too small to generate an ction potential. (When action potentials are elicted in visceral smooth muscle, as many as 30 to 0 smooth muscle fibers must depolarize simultaeously before a self-propagating action potential nsues.) Yet, even without an action potential in e multiunit smooth muscle fibers, the local deolarization caused by the nerve transmitter subitself spreads "electrotonically" over the ntire fiber and is all that is needed to cause the uscle contraction.

XCITATION-CONTRACTION COUPLING—OLE OF CALCIUM IONS

In the previous chapter it was pointed out that he actual contractile process in skeletal muscle ia ctivated by calcium ions. This is also true in mooth muscle. However, the source of the calcium ons differs in smooth muscle because the sarcolasmic reticulum of smooth muscle is poorly deeloped.

In some types of smooth muscle, almost all the alcium ions that cause contraction enter the musle cell from the extracellular fluid at the time of he action potential. There is a reasonably higb oncentration of calcium ions in the extracellular luid, greater than 10^{-3} molar in comparison with less than 10^{-7} molar in the cell, and as was pointed out in the previous section, the smooth muscle action potential is caused at least partly by influx of calcium ions into the muscle fiber. Because the smooth muscle fibers are extremely small (in contrast to the sizes of the skeletal muscle fibers), these calcium ions can diffuse to all parts of the smooth muscle and elicit the contractile process. The time required for this diffusion to occur is usually 200 to 300 milliseconds and is called the *latent period* before the contraction begins; this latent period is some 50 times as great as that for skeletal muscle contraction.

Yet, in some smooth muscle there is a moderately developed sarcoplasmic reticulum. However, there are no T tubules. Instead, the cisternae of the reticulum abut the cell membrane. Therefore, it is believed that the membrane action potentials. in these smooth muscle fibers cause release of calcium ions from these cisternae, thereby providing a greater degree of contraction than would occur on the basis of calcium ions entering through the cell membrane alone.

The Calcium Pump. To cause relaxation of the smooth muscle contractile elements, it ia necessary to remove the calcium ions. This removal ia achieved by a calcium pump that pumps the calcium ions out of the smooth muscle fiber and back into the extracellular fluid, or pumps the calcium iona into the sarcoplasmic reticulum. However, this pump is very slow acting in comparison with the fast-acting sarcoplasmic reticulum pump in skeletal muscle. Therefore, the duration of smooth muscle contraction is often in the order of seconds rather than in tens of milliseconds, as occurs for skeletol muscle.

Mechanism by Which Calcium Iona Excite Contraction in Smooth Muscle. The mechanism by which calcium ions excite contraction in smooth muscle fibers is quite different from that in skeletal muscle. In fact, smooth muscle probably does not have an effective troponin complex, the factor that is activated by calcium ions in skeletal muscles to initiate contraction. Instead, the calcium ions excite contraction in smooth muscle fibers by activating the ATPase activity of the myosin heads. When calcium is not present, this ATPase activity in smooth muscle is extremely slight so that ATP cannot be cleaved and the contractile process cannot take place. Calcium ions activate this ATPase activity in the following way:

The first event is the increase in calcium ion concentration as a result of the action potential or of any other stimulus that causes entry of calcium ions into the cell cytosol. Second, the calcium ions bind with a protein, *calmodulin,* that is very similar to the troponin C in skeletal muscle fibers. Third, the bound product of calmodulin and calcium ions in turn binds with or activates one of the light chain polypeptides of the myosin head; this in turn activates the ATPase activity of the

myosin head. Fourth, the newly excited ATPase activity of the head then causes cleavage of ATP and the usual conformational changes in the head, leading to the same "walk-along" contractile process in smooth muscle fibers as occurs in skeletal muscle.

However, two other features of this contractile activating process are also considerably different from the activation of skeletal muscle First, all these events are very slow to take place, which further explains the slowness and long duration of smooth muscle contraction. Second, other factors besides calcium ions can have large effects on the intensity of the contractile process, which usually is not true in skeletal muscle. One of these is the concentration of cyclic AMP in the cell, which, we shall see later, is another means for influencing the strength of smooth muscle contraction.

NEUROMUSCULAR JUNCTIONS OF SMOOTH MUSCLE

Physiologic Anatomy of Smooth Muscle Neuromuscular Junctions. Neuromuscular junctions of the type found on skeletal muscle fibers do not occur in smooth muscle. Instead, the nerve fibers generally branch diffusely on top of a sheet of muscle fibers, as illustrated in Figure 12–7. In most instances these fibers do not make direct contact with the smooth muscle fibers at all but instead form so-called *diffuse junctions* that secrete their transmitter substance into the interstitial fluid from a few nanometers to a few microns away from the muscle cells; the transmitter substance then diffuses to the cells. Furthermore, where there are many layers of muscle cells, the nerve fibers often innervate only the outer layer, and the muscle excitation then travels from this outer layer to the inner layers by direct action potential conduction or by subsequent diffusion of the transmitter substance But, less often, terminal branches of the axons do penetrate into the muscle mass.

The axons innervating smooth muscle fibers also do not have typical branching end-feet, as observed in the end-plate on skeletal muscle fibers. Instead most of the fine terminal axons have multiple varicosities spread along their axes. At these points the Schwann cells are interrupted so that transmitter substance can be secreted through the walls of the varicosities. In the varicosities are vesicles similar to those present in the skeletal muscle end-plate containing transmitter substance However, in contrast to the vesicles of skeletal muscle junctions that contain only acetylcholine, the vesicles of the autonomic nerve fiber varicosities contain acetylcholine in some fibers and norepinephrine in others.

In a few instances, particularly in the multiunit type of smooth muscle, the varicosities lie directly on the muscle fiber membrane with a separation from this membrane of as little as 20 nanometers—the same width as the synaptic cleft that occurs in the skeletal muscle junction These *contact junctions* function in much the same way as the skeletal muscle neuromuscular junction, and the latent period of contraction of these smooth muscle fibers is considerably shorter than of fibers stimulated by the diffuse junctions.

Excitatory and Inhibitory Transmitter Substances at the Smooth Muscle Neuromuscular Junction. Two different transmitter substances known to be secreted by the autonomic nerves innervating smooth muscle are *acetylcholine* and *norepinephrine*. Acetylcholine is an excitatory transmitter substance for smooth muscle fibers in some organs but an inhibitory substance for smooth muscle in other organs And when acetylcholine excites a muscle fiber, norepinephrine ordinarily inhibits it. Or when acetylcholine inhibits a fiber, norepinephrine usually excites it.

But why these different responses? The answer is that both acetylcholine and norepinephrine excite or inhibit smooth muscle by first binding with a *receptor protein* on the surface of the muscle cell membrane This receptor in turn controls the opening or closing of ion channels or controls some other means for activating or inhibiting the smooth muscle fiber. Furthermore, some of the receptor proteins are *excitatory receptors* whereas others are *inhibitory receptors*. Thus, it is the type of receptor that determines whether the smooth muscle will be inhibited or excited and also determines which of the two transmitters, acetylcholine or norepinephrine, will be effective in causing the excitation or inhibition. These receptors will be discussed in more detail in Chapter 57 in relation to the function of the autonomic nervous system

Excitation of Action Potentials in Smooth Muscle Fibers—The Junctional Potential. Transmission of impulses from terminal nerve fibers to smooth muscle fibers occurs in very much the same manner as transmission at the neuromuscular junction of skeletal muscle fibers except for temporal differences. When an action potential reaches the terminal of an excitatory nerve fibril,

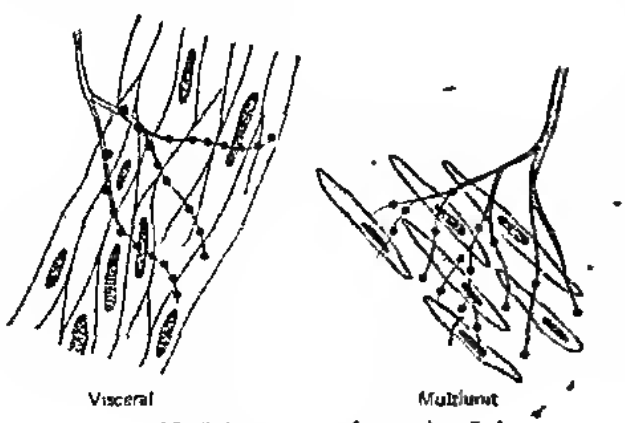


Figure 12–7. Innervation of smooth muscle

here is a typical latent period of 50 milliseconds before any change in the membrane potential of the smooth muscle fiber can be detected. Then the potential rises to a maximal level in approximately 200 milliseconds. If an action potential does not occur, this potential gradually disappears at a rate of approximately one-half every 200 to 500 milliseconds. This complete sequence of potential changes is called the *junctional potential;* it is analogous to the end-plate potential of the skeletal muscle fibers except that its duration is 20 to 100 times as long.

If the junctional potential rises to the threshold level for discharge of the smooth muscle membrane, an action potential will occur in the smooth muscle fiber in exactly the same way that an action potential occurs in a skeletal muscle fiber. A typical smooth muscle fiber has a normal resting membrane potential of −50 to −60 millivolts, and the threshold potential at which the action potential occurs is about −30 to −35 millivolts.

Inhibition at the Smooth Muscle Neuromuscular Junction. When a transmitter substance at the nerve ending interacts with an inhibitory receptor instead of an excitatory receptor, the membrane potential of the muscle fiber becomes more negative than ever, for instance down to −70 millivolts; that is, it becomes *hyperpolarized* and therefore becomes much more difficult to excite than is usually the case.

SMOOTH MUSCLE CONTRACTION WITHOUT ACTION POTENTIALS—EFFECT OF LOCAL TISSUE FACTORS AND HORMONES

Though we have thus far discussed smooth muscle contraction elicited only by nervous signals and smooth muscle membrane action potentials, we must quickly disavow the fact that all smooth muscle contraction occurs in this way. In fact, probably half or more of all smooth muscle contraction is initiated not by action potentials but by stimulatory factors acting directly on the smooth muscle contractile machinery. The two types of non-nervous and nonaction potential stimulating factors most often involved are (1) local tissue factors, and (2) various hormones.

Smooth Muscle Contraction in Response to Local Tissue Factors. In Chapter 20 we shall discuss the control of contraction of the arterioles, meta-arterioles, and precapillary sphincters. The smaller of these vessels have little or no nervous supply. Yet, the smooth muscle is highly contractile, responding rapidly to changes in local conditions in the surrounding interstitial fluid. In this way a powerful local feedback control system controls the blood flow to the local tissue area. Some of the specific control factors are:

(1) Lack of oxygen in the local tissues causes smooth muscle relaxation and therefore vasodilatation.

(2) Excess carbon dioxide causes vasodilatation.

(3) Increased hydrogen ion concentration also causes increased vasodilatation.

And such factors as lactic acid, increased potassium ions, diminished calcium ion concentration, and decreased body temperature will also cause local vasodilatation.

Effects of Hormones on Smooth Muscle Contraction. Most of the circulating hormones in the body affect smooth muscle contraction at least to some degree, and some have very profound effects. Some of the more important blood-borne hormones are norepinephrine, epinephrine, acetylcholine, angiotensin, vasopressin, oxytocin, serotonin, and histamine.

A hormone will cause contraction of smooth muscle when the smooth muscle cells contain an *excitatory receptor* for the respective hormone. However, the hormone will cause inhibition instead of contraction if the cells contain an *inhibitory receptor* rather than an excitatory receptor. Thus, most of the hormones will cause excitation of some smooth muscle but inhibition of other muscle

Some of the hormones—especially norepinephrine, vasopressin, and angiotensin—have such a powerful excitatory effect that they can cause smooth muscle spasm for hours.

Mechanism of Smooth Muscle Excitation by Local Tissue Factors and Hormones. Most smooth muscle contraction caused by hormones results from an initial increase in calcium ion concentration in the muscle cell, as occurs when an action potential excites the fiber. Each hormone usually binds with a specific receptor protein in the membrane and initiates the contraction, usually by opening calcium ion channels and decreasing the membrane potential a moderate amount—that is, "depolarizing" the membrane—but usually *without causing an action potential.* Nevertheless, the flow of calcium ions to the interior of the cell causes contraction.

In some instances, however, smooth muscle contraction is initiated by hormones without a change in the membrane potential at all. In these instances it is believed that the exciting hormone binds with a receptor protein that activates a nonvoltage-related mechanism that leads to contraction. One such mechanism is activation of the enzyme *adenylcyclase* in the cell membrane; a portion of this enzyme that protrudes to the interior of the cell then causes the formation of *cyclic AMP,* which is a so-called *second messenger.* The cyclic AMP in turn can promote calcium ion release from intracellular organelles, such as the limited sarcoplasmic reticulum or even the mitochondria. In addition, as noted earlier, the cyclic AMP itself has the direct effect of increasing the ATPase activity of the myosin heads and can in this way also increase the degree of smooth muscle contraction.

Unfortunately, it is not known how most other local tissue factors besides the hormones, such as oxygen lack, excess CO_2, or changes in hydrogen ion concentration, either excite or inhibit smooth muscle contraction. However, possible mechanisms include changes in the cell membrane potential, changes in the permeability of the membrane to calcium ions, and/or changes in the intracellular contractile machinery.

MECHANICAL CHARACTERISTICS OF SMOOTH MUSCLE CONTRACTION

From the foregoing discussion of the many different types of smooth muscle and the different ways in which contraction can be elicited, one can readily understand why smooth muscle in different parts of the body has many different characteristics of contraction. For instance, the multiunit smooth muscle of the large blood vessels contracts mainly in response to nerve impulses, whereas in many types of visceral smooth muscle—the smaller blood vessels, the ureter, the bile ducts, and other glandular ducts—a self-excitatory process controlled mainly by local factors and by hormones causes continuous rhythmic contraction.

Tone of Smooth Muscle. Smooth muscle can maintain a state of long-term, steady contraction that has been called either tonus contraction of smooth muscle or simply *smooth muscle tone*. This is an important feature of smooth muscle contraction because it allows prolonged or even indefinite continuance of the smooth muscle function. For instance, the arterioles are maintained in a state of tonic contraction almost throughout the entire life of the person. Likewise, tonic contraction in the gut wall maintains steady pressure on the contents of the gut, and tonic contraction of the urinary bladder wall maintains a moderate amount of pressure on the urine in the bladder.

Tonic contractions of smooth muscle can be caused in either of two ways.

(1) They are sometimes caused by *summation of individual contractile pulses*, each contractile pulse is initiated by a separate action potential in the same way that tetanic contractions are produced in skeletal muscle.

(2) However, most smooth muscle tonic contractions probably result from *prolonged direct smooth muscle excitation* without action potentials, usually caused by local tissue factors, or circulating hormones. For instance, prolonged tonic contractions of the blood vessels without the mediation of action potentials are regularly caused by angiotensin, vasopressin, or norepinephrine, and these play an important role in the long-term regulation of arterial pressure, as will be discussed in Chapters 21 and 22.

A typical example of a tonic contraction is illustrated in Figure 12–8. This figure also shows that rhythmic contractions can be superimposed onto the tonic contraction because of simultaneous (a) rhythmical nerve discharges or (b) rhythmical pacemaker slow waves that periodically excite the smooth muscle.

Degree of Shortening of Smooth Muscle During Contraction. A special characteristic of smooth muscle—one that is also different from skeletal muscle—is its ability to shorten a far greater percentage of its length than can skeletal muscle. Skeletal muscle has a

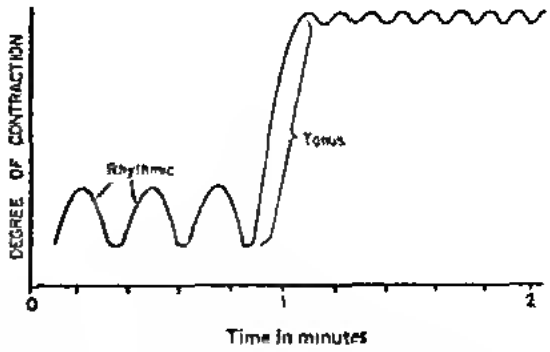


Figure 12–8. Record of rhythmic and tonic smooth m contraction.

useful distance of contraction equal to only 25 to 35 cent of its length, while smooth muscle can often tract quite effectively to as short as one half or even fourth its stretched length, giving as much as a four distance of contraction. This allows smooth muscle perform important functions in the hollow viscera– instance, allowing the gut, the bladder, blood vess and other internal structures of the body to change t lumen diameters from almost zero to very large val

Stress-Relaxation of Smooth Muscle. Another portant characteristic of smooth muscle is its sbilit change length greatly without marked changes in sion. This results from a phenomenon called *str relaxation*, which may be explained as follows:

If a segment of smooth muscle 1 inch long is sudde stretched to 2 inches, the tension between the two e increases tremendously at first, but the extra tens also begins to disappear immediately, and within s minutes it has returned almost to its level prior to stretch, even though the muscle is now twice as lo This possibly results from the loose arrangement of actin and myosin filaments in smooth muscle. Ove period of time, the filaments of the stretched mu presumably rearrange their bonds and gradually al the sliding process to take place, thus allowing tension to return almost to its original amount.

Exactly the converse effect occurs when smooth m ele is shortened. Thus, if the 2 inch segment of smo muscle is shortened back to 1 inch, essentially tension will be lost from the muscle immediately. Gr ually, over a period of 1 minute or more, much of tension returns, this again presumably resulting fr alow sliding of the filaments. This is called *reverse str relaxation*.

REFERENCES

Bennett, M. R. Development of neuromuscular synapses. *Physiol* 63:915, 1983.

Blaustein, M. P., and Hamlyn, J. M. Sodium transport inhibition, calcium, and hypertension. The natriuretic hormone/Na+=Ca exchange/hypertension hypothesis. *Am. J. Med.*, 77:45, 1984.

Bohr, D. F., and Webb, R. C. Vascular smooth muscle function an changes in hypertension. *Am. J. Med.*, 77:3, 1984.

Bolton, T. B. Mechanisms of action of transmitters and other substa on smooth muscle. *Physiol. Rev.*, 59:606, 1979.

Borgstrom, P., *et al.* An evaluation of the metabolic interaction myogenic vascular reactivity during blood flow autoregulation. *Physiol. Scand.*, 122:275, 1984.

Bulbring, E. (ed.). Smooth Muscle: An Assessment of Current Kn edge. Austin, University of Texas Press, 1981.

Bulbring, E., *et al.* (eds.). Physiology of Smooth Muscles. Twenty-s International Congress of Physiological Sciences. New York, Ra Press, 1975.

Beccarelli, B, and Hurlbut, W P Vesicle hypothesis of the release of quanta of acetylcholine *Physiol Rev*, *60* 396, 1980
Chamley-Campbell, J, et al Smooth muscle cell in culture *Physiol Rev*, *59* 1, 1979
Changeux, J P, et al Acetylcholine receptor An allosteric protein *Science*, *225* 1335, 1984
Cotman, C W, et al Synapse replacement in the nervous system of adult vertebrates. *Physiol Rev*, *61* 684 1981
Daniel, E E, and Sarna, S The generation and conduction of activity in smooth muscle *Annu Rev Pharmacol Toxicol*, *18* 145, 1978
Dau, P C (ed) Plasmapheresis and the Immunobiology of Myasthenia Gravis Boston, Houghton Mifflin, 1979
Dowben, R M (ed) Cell and Muscle Motility New York, Plenum Publishing Corp, 1983
Duckles, S P, and Banner, W, Jr Changes in vascular smooth muscle reactivity during development *Annu Rev Pharmacol Toxicol*, *24* 65, 1984
Eisenberg, E, and Greene, L E The relation of muscle biochemistry to muscle physiology *Annu Rev Physiol*, *42* 293 1980
Fambrough, D M Control of acetylcholine receptors in skeletal muscle *Physiol. Rev*, *59* 165, 1979
Fleming, W, W The electrogenic Na⁺,K⁺ pump in smooth muscle Physiologic and pharmacologic significance *Annu Rev Pharmacol Toxicol*, *20* 129, 1980
Furchgott, R F The role of endothelium in the responses of vascular smooth muscle to drugs *Annu Rev Pharmacol Toxicol* *24* 175, 1984
Gabella, G; Structural apparatus for force transmission in smooth muscle *Physiol Rev*, *64* 455, 1984
Geige, P W; Generation of end plate potentials. *Physiol Rev* *56* 177 1976
Guyton, A C, and MacDonald, M A Physiology of botulinus toxin *Arch Neurol Psychiat*, *57* 578, 1947
Guyton, A C, and Reeder, R C The dynamics of curarization *J Pharmacol Exp Ther.*, *97* 322, 1949
Hartshorne, D J., and Gorecka, A Biochemistry of the contractile proteins of smooth muscle *In* Bohr D F, *et al (eds)* Handbook of Physiology Sec. 2, Vol II Baltimore, Williams & Wilkins, 1980 p 83
Hartshorne, D J, and Siemankowski R F Regulation of smooth muscle actomyosin *Annu Rev Physiol* *43* 319, 1981
Hertzberg, E L., et al Gap junctional communication *Annu Rev Physiol*, *43* 479, 1981
Hess, G. P, et al Acetylcholine receptor-controlled ion translocation Chemical kinetic investigations of the mechanism *Annu Rev Biophys Bioeng*, *12* 443, 1983
Johansson, B. Vascular smooth muscle reactivity *Annu Rev Physiol* *43* 359, 1981.
Johansson, B, and Somlyo, A. P Electrophysiology end excitation-contraction coupling *In* Bohr, D F, *et al* (eds) Handbook of Physiology Sec 2, Vol II Baltimore, Williams & Wilkins, 1980, p. 301
Klein, W L Biochemistry and regulation of signal transduction by neuronal acetylcholine receptors *Curr Top Cell Regul*, *24*, 129, 1984
Lambert, J J, et al Drug-induced modification of ionic conductance at the neuromuscular junction *Annu Rev Pharmacol Toxicol*, *23* 505, 1983
Landmesser, L. T The generation of neuromuscular specificity *Annu Rev Neurosci*, *3* 279, 1980
Lindstrom, J, and Dau, P Biology of myasthenia gravis *Annu Rev Pharmacol Toxicol* *20* 337, 1980
Loewenstein, W R. Junctional intercellular communication The cell-to-cell membrane channel *Physiol Rev*, *61* 829, 1981
Loh, H H, and Law, P Y The role of membrane lipids in receptor mechanisms *Annu Rev Pharmacol Toxicol*, *20* 201, 1980
Mark, R F Synaptic repression at neuromuscular junctions *Physiol Rev*, *60* 355, 1980
McKinney, M, and Richelson, E The coupling of neuronal muscarinic receptor to responses *Annu Rev Pharmacol Toxicol* *24* 121, 1984
Morgan, D L, and Proske, U Vertebrate slow muscle Its structure, pattern of innervation, and mechanical properties *Physiol Rev*, *64* 103, 1984
Murphy, R A Filament organization and contractile function in vertebrate smooth muscle *Annu Rev Physiol*, *41* 737, 1979
Paul, R J Chemical energetics of vascular smooth muscle *In* Bohr, D F, *et al* (eds) Handbook of Physiology Sec 2, Vol II Baltimore, Williams & Wilkins, 1980, p 201
Peper K, et al The acetylcholine receptor et the neuromuscular junction *Physiol Rev* *62* 1271, 1982
Popot, J-L, and Changeux, J P Nicotinic receptor of acetylcholine Structure of an oligomeric integral membrane protein *Physiol Rev*, *64* 1162, 1984
Prosser, C L Evolution and diversity of nonstriated muscles *In* Bohr, D F, *et al* (eds) Handbook of Physiology Sec 2, Vol II Baltimore, Williams & Wilkins, 1980, p 635
Pumplin D W, and Fambrough, D M Turnover of acetylcholine receptors in skeletal muscle *Annu Rev Physiol*, *44* 319, 1982
Purves, D, and Lichtman, J W Specific connections between nerve cells *Annu Rev Physiol*, *45* 553, 1983
Schneider, M F Membrane charge movement and depolarization-contraction coupling *Annu Rev Physiol* *43* 507, 1981
Somlyo, A P Ultrastructure of vascular smooth muscle *In* Bohr, D F, *et al* (eds) Handbook of Physiology Sec 2, Vol II Baltimore, Williams & Wilkins, 1980, p 33
Taugt, L Nonvascular release of neurotransmitter *Physiol Rev*, *62* 857, 1982
Venhoutte, P M Calcium-entry blockers, vascular smooth muscle and systemic hypertension *Am J Cardiol* *55* 17B, 1985

IV

THE HEART

13 ■ Heart Muscle; The Heart as a Pump

14 ■ Rhythmic Excitation of the Heart

15 ■ The Normal Electrocardiogram

16 ■ Electrocardiographic Interpretation in Cardiac Muscle Abnormalities—Vectorial Analysis

17 ■ Electrocardiographic Interpretation of Cardiac Arrhythmias

13

Heart Muscle; The Heart As a Pump

With this chapter we begin discussion of the heart and circulatory system The heart is actually two separate pumps, a *right heart* that pumps the blood through the lungs and a *left heart* that pumps the blood through the peripheral organs. In turn, each of these two separate hearts is a pulsatile two-chamber pump composed of an *atrium* and a *ventricle*. Tha atrium functions principally as an entryway to the ventricle, but it also pumps weakly to help move the blood into the ventricle. The ventricle in turn supplies the main force that propels the blood through either the pulmonary or the peripheral circulation.

Special mechanisms in the heart maintain cardiac rhythmicity and transmit action potentials throughout the heart muscle to cause the heart's rhythmical beat. This rhythmical control system will be explained in the following chapter In the present chapter we will explain how the heart operates as a pump—that is, the function of the heart muscle, of the valves, of the various chambers of the heart, and, especially important, how the pumping activity of the heart is regulated so that the heart can pump either small or large amounts of blood as dictated by the needs of the body

PHYSIOLOGY OF CARDIAC MUSCLE

The heart is composed of three major types of cardiac muscle: atrial muscle, ventricular muscle, and specialized excitatory and conductive muscle fibers The atrial and ventricular types of muscle contract in much the same way as skeletal muscle, except that the duration of contraction is much longer On the other hand, the specialized excitatory and conductive fibers contract only feebly because they contain few contractile fibrils; instead, because of their rhythmical properties and their rapidity of conduction, they provide an excitatory system for the heart and a transmission system for rapid conduction of the cardiac excitatory signal throughout the heart.

PHYSIOLOGIC ANATOMY OF CARDIAC MUSCLE

Figure 13–1 illustrates a typical histologic picture cardiac muscle, showing the cardiac muscle fibers ranged in a latticework, the fibers dividing, then rec bining, and then spreading again. One notes imme ately from this figure that cardiac muscle is *striated* the same manner as typical skeletal muscle. Furth more, cardiac muscle has typical myofibrils that cont *actin* and *myosin filaments* almost identical to th found in skeletal muscle, and these filaments interd tate and slide along each other during the process contraction in the same manner as occurs in skele muscle (See Chapter 11.)

Cardiac Muscle as a Syncytium. The anguis dark areas crossing the cardiac muscle fibers in Fig 13–1 are called *intercalated discs;* however, they actually cell membranes that separate individual c diac muscle cells from each other That is, cardiac mus fibers are made up of many cardiac muscle cells c nected in series with each other. Yet electrical resista through the intercalated disc is only one four-hundre the resistance through the outside membrane of cardiac muscle fiber, because the cell membranes f with each other and form very permeable junctions t allow relatively free diffusion of ions. Therefore, from functional point of view, ions move with ease along axes of the cardiac muscle fibers so that action potenti

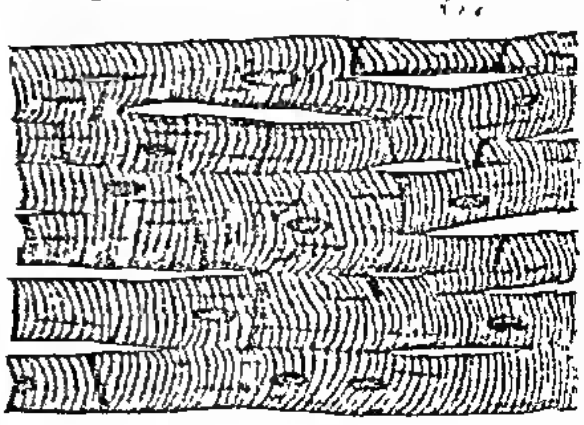


Figure 13–1. The "syncytial," interconnecting nature of cardiac muscle.

travel from one cardiac muscle cell to another, past the intercalated discs, with only slight hindrance. Therefore, cardiac muscle is a *syncytium*, in which the cardiac muscle cells are so tightly bound that when one of these cells becomes excited, the action potential spreads to all of them, spreading from cell to cell and spreading throughout the latticework interconnections.

The heart is composed of two separate syncytiums, the *atrial syncytium* that constitutes the walls of the two atria and the *ventricular syncytium* that constitutes the walls of the two ventricles. These are separated from each other by fibrous tissue that surrounds the valvular openings between the atria and ventricles, but action potentials can be conducted from the atrial syncytium into the ventricular syncytium by way of a specialized conductive system, the *A-V bundle*, which will be discussed in detail in the following chapter. This division of the muscle mass of the heart into two separate functional syncytia allows the atria to contract a short time ahead of ventricular contraction, which is important for the effectiveness of heart pumping

All-or-Nothing Principle As Applied to the Heart. Because of the syncytial and interconnecting nature of cardiac muscle, stimulation of any single atrial muscle fiber causes the action potential to travel over the entire atrial muscle mass; and, similarly, stimulation of any single ventricular fiber causes excitation of the entire ventricular muscle mass. If the A-V bundle is intact, the action potential passes also from the atria to the ventricles. This is called the all-or-nothing principle, and it is precisely the same as that discussed in Chapter 10 for nerve fibers. However, because the cardiac muscle fibers all interconnect with each other, the all-or-nothing principle applies to the entire functional syncytium of the heart rather than to single muscle fibers, as in the case of skeletal muscle fibers.

ACTION POTENTIALS IN CARDIAC MUSCLE

The *resting membrane potential* of normal cardiac muscle is approximately −85 to −95 millivolt (mv) and approximately −90 to −100 mv in the specialized conductive fibers, the Purkinje fibers, which are discussed in the following chapter.

The *action potential* recorded in ventricular muscle, shown by the bottom record of Figure 13–2, is 105 mv, which means that the membrane potential rises from its normally very negative value to a slightly positive value of about +20 mv. Because this change of potential from negative to positive, the positive portion is called the *overshoot potential*. Then, after the initial *spike*, the membrane remains depolarized for about 0.2 second in atrial muscle and about 0.3 second in ventricular muscle, exhibiting a *plateau* as illustrated in Figure 13–2, followed at the end of the plateau by abrupt repolarization. The presence of this plateau in the action potential causes muscle contraction to last 20 to 50 times as long in cardiac muscle as in skeletal muscle.

At this point we must ask the question: Why is the action potential of cardiac muscle so long and why does it have a plateau while that of skeletal

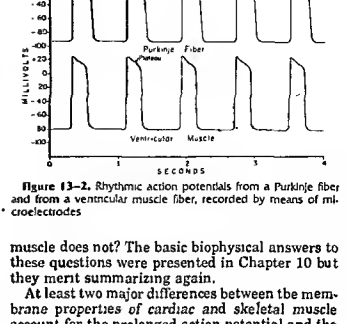


Figure 13–2. Rhythmic action potentials from a Purkinje fiber and from a ventricular muscle fiber, recorded by means of microelectrodes

muscle does not? The basic biophysical answers to these questions were presented in Chapter 10 but they merit summarizing again.

At least two major differences between the membrane properties of cardiac and skeletal muscle account for the prolonged action potential and the plateau in cardiac muscle.

First, the action potential of skeletal muscle is caused almost entirely by sudden opening of large numbers of so-called *fast sodium channels* that allow tremendous numbers of sodium ions to enter the skeletal muscle fiber. These channels are called "fast" channels because they remain open for only a few ten-thousandths of a second and then abruptly close. At the end of this closure, the process of repolarization occurs, and the action potential is over within another ten-thousandth of a second or so. In cardiac muscle, on the other hand, the action potential is caused by the opening of two types of channels, (1) the same *fast sodium channels* as those in skeletal muscle, and (2) another entire population of so-called *slow calcium-sodium channels*. This second population of channels differs from the fast sodium channels in being slower to open but, more importantly, they remain open for several tenths of a second. During this time large amounts of both calcium and sodium ions continue to flow through the channels to the interior of the cardiac muscle fiber, and this maintains a prolonged period of depolarization. It is this that causes the plateau in the action potential. Furthermore, the calcium ions that enter the muscle during this action potential play an important role in helping excite the muscle contractile process, which is another difference between cardiac muscle and skeletal muscle that we shall discuss later in this chapter.

The second major functional difference between cardiac muscle and skeletal muscle that helps account for both the prolonged action potential and its plateau is this: Immediately after the onset of the action potential, the permeability of the car-

diac muscle membrane for potassium *decreases* about fivefold, an effect that does not occur in skeletal muscle It is possible that this decreased potassium permeability is caused in some way by the excess calcium influx through the calcium channels just noted. But, regardless of the cause, the decreased potassium permeability greatly decreases the outflux of potassium ions during the action potential plateau and thereby prevents early recovery. When the slow calcium-sodium channels close at the end of 0.2 to 0.3 second and the influx of calcium and sodium ions ceases, then the membrane permeability for potassium increases very rapidly, and the rapid loss of potassium from the fiber returns the membrane potential to its resting level, thus ending the action potential.

Velocity of Conduction in Cardiac Muscle. The velocity of conduction of the action potential in both atrial and ventricular muscle fibers is about 0.3 to 0.5 meter per second, or about 1/250 the velocity in very large nerve fibers and about one tenth the velocity in skeletal muscle fibers. The velocity of conduction in the specialized conductive system varies from 0.02 to 4 meters per second in different parts of the system, as is explained in the following chapter

Refractory Period of Cardiac Muscle. Cardiac muscle, like all excitable tissue, is refractory to restimulation during the action potential An extremely strong electrical stimulus can sometimes initiate a new spike at the very end of the action potential plateau, but this is a very abnormal situation and the spike is not propagated along the muscle. Therefore, the refractory period of the heart is usually stated to be the interval of time, as shown to the left in Figure 13–3, during which a normal cardiac impulse cannot re-excite an already excited area of cardiac muscle. The normal refractory period of the ventricle is 0.25 to 0.3 second, which is approximately the duration of the action potential. There is an additional *relative refractory period* of about 0.05 second during whic the muscle is more difficult than normal to exci but nevertheless can be excited, as illustrated l the early premature contraction in Figure 13–3

The refractory period of atrial muscle is muc shorter than that for the ventricles (about 0 1 second), and the relative refractory period is a other 0 03 second. Therefore, the rhythmical ra of contraction of the atria can be much faster tha that of the ventricles.

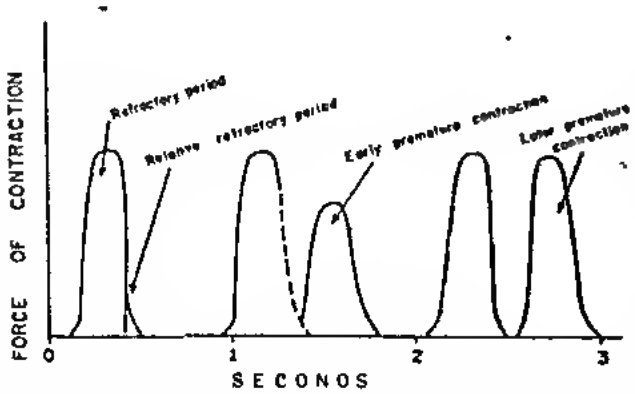


Figure 13–3. Contraction of the heart, showing the durations of the refractory period and the relative refractory period, the effect of an early premature contraction, and the effect of a later premature contraction. Note that the premature contractions do not cause wave summation as occurs in skeletal muscle

CONTRACTION OF CARDIAC MUSCLE

Excitation-Contraction Coupling—Functi of Calcium Ions and of the T Tubules. The ter excitation-contraction coupling means the mech nism by which the action potential causes t myofibrils of muscle to contract. This was discuss for skeletal muscle in Chapter 11. However, on again there are differences in this mechanism cardiac muscle that have important effects on t characteristics of cardiac muscle contraction.

As is true for skeletal muscle, whea an acti potential passes over the cardiac muscle mer brane, the action potential also spreads to t interior of the cardiac muscle fiber along the mer branes of the T tubules. The T tubule acti potentials in turn cause instantaneous release calcium ions into the muscle sarcoplasm from t cisternae of the sarcoplasmic reticulum. Then t calcium ions diffuse in another few thousandths a second into the myofibrils where they catalyz the chemical reactions that promote sliding of t actin and myosin filaments along each other; th in turn produces the muscle contraction.

Thus far, this mechanism of excitation-contra tion coupling is the same as that for skelet muscle, but at this point a major difference begi to appear. In addition to the calcium ions that a released into the sarcoplasm from the cisternae the sarcoplasmic reticulum, large quantities calcium ions also diffuse into the sarcoplasm fro the T tubules at the time of the action potentia Indeed, without this extra calcium from the tubules, the strength of cardiac muscle contracti would be considerably reduced, because the ciste nae of cardiac muscle are less well developed th those of skeletal muscle and do not store enou calcium. On the other hand, the T tubules cardiac muscle have a diameter five times as gre as that of the skeletal muscle tubules and a volum 25 times as great; also, inside the T tubules is large quantity of mucopolysaccharides that a electronegatively charged and bind an abunda store of calcium ions, keeping this always availab for diffusion to the interior of the cardiac musc fiber when the T tubule action potential occurs

The strength of contraction of cardiac depends to a great extent on the concentration calcium ions in the extracellular fluids. The re for this is that the ends of the T tubules op

irectly to the outside of the cardiac muscle fibers, llowing the same extracellular fluid that is in the ardiac muscle interstitium to percolate through he T tubules as well. Consequently, the quantity f calcium ions in the T tubule system, as well as h availability of calcium ions to cause cardiac contraction, depends directly on the extraellular fluid calcium ion concentration.

By way of contrast, the strength of skeletal nuscle contraction is hardly affected by the extraellular fluid calcium concentration because its ontraction is caused almost entirely by calcium ons released from the sarcoplasmic reticulum inside the skeletal muscle fiber itself.

At the end of the plateau of the action potential he influx of calcium ions to the interior of the fiber is suddenly cut off, and the calcium ons in the sarcoplasm are rapidly pumped back nto both the sarcoplasmic reticulum and the T ubules. As a result, the contraction ceases until a new action potential occurs.

Another difference between the tubules of cardiac nuscle and those of skeletal muscle is that of the T ubules in the skeletal muscle are located adjacent to the actin and myosin filaments where they overlap each other so that there are two T tubule systems to each sarcomere. On the other hand, in cardiac muscle there is only one T tubule system per sarcomere located at the Z line but composed of much larger T tubules. This difference in structure is compatible with the fact that cardiac muscle contracts much more slowly than skeletal muscle so that there is adequate time for calcium ions to diffuse from the Z line to the middle of the sarcomere where the contractile process occurs.

Duration of Contraction. Cardiac muscle begins to contract a few milliseconds after the action potential begins and continues to contract for a few milliseconds after the action potential ends. Therefore, the duration of contraction of cardiac muscle is mainly a function of the duration of the action potential—about 0.2 second in atrial muscle and 0.3 second in ventricular muscle.

Effect of Heart Rate on Duration of Contraction. When the heart rate increases, the duration of each total cycle of the heart, including both the contraction phase and the relaxation phase, obviously decreases. The duration of the action potential and the period of contraction (systole) also decrease but not as great a percentage as does the relaxation phase (diastole). At a normal heart rate of 72 beats per minute, the period of contraction is about 0.40 of the entire cycle. At 3 times normal heart rate, this period is about 0.65 of the entire cycle, which means that the heart under some conditions does not remain relaxed long enough to allow complete filling of the cardiac chambers prior to the next contraction.

THE CARDIAC CYCLE

Figure 13–4 illustrates the physical structure of the heart and also the course of blood flow through the heart. It is also clear that the heart is in reality four separate pumps: two *primer pumps*, the *atria*, and two *power pumps*, the *ventricles*.

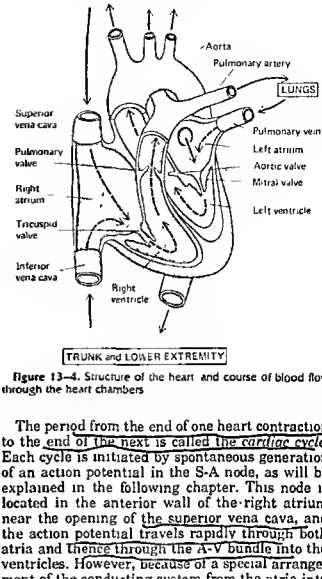


Figure 13–4. Structure of the heart and course of blood flow through the heart chambers

The period from the end of one heart contraction to the end of the next is called the *cardiac cycle*. Each cycle is initiated by spontaneous generation of an action potential in the S-A node, as will be explained in the following chapter. This node is located in the anterior wall of the right atrium near the opening of the superior vena cava, and the action potential travels rapidly through both atria and thence through the A-V bundle into the ventricles. However, because of a special arrangement of the conducting system from the atria into the ventricles, there is a delay of more than 1/10 second between passage of the cardiac impulse from the atria into the ventricles. This allows the atria to contract ahead of the ventricles, thereby pumping blood into the ventricles prior to the very strong ventricular contraction. Thus, the atria act as primer pumps for the ventricles, and the ventricles then provide the major source of power for moving blood through the vascular system.

SYSTOLE AND DIASTOLE

The cardiac cycle consists of a period of relaxation called *diastole* followed by a period of contraction called *systole*.

Figure 13–5 illustrates the different events during the cardiac cycle. The top three curves show the pressure changes in the aorta, the left ventricle, and the left atrium, respectively. The fourth curve depicts the changes in ventricular volume, the fifth the electrocardiogram, and the sixth a

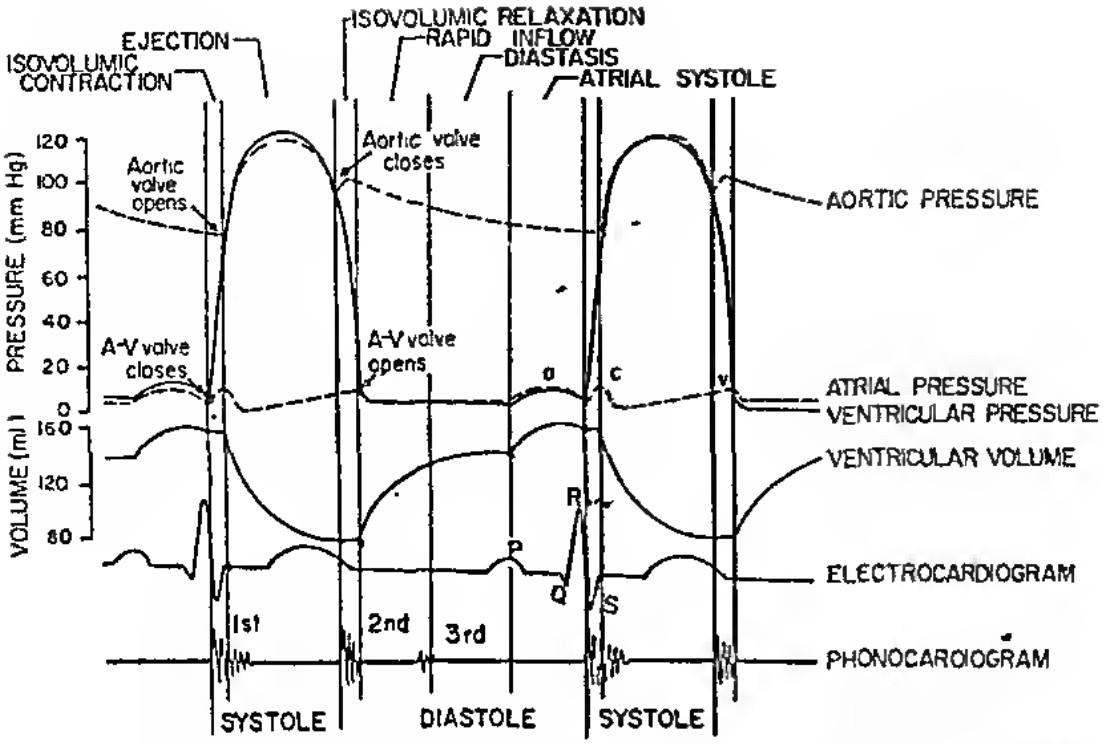


Figure 13–5. The events of the cardiac cycle, showing changes in left atrial pressure, left ventricular pressure, aortic pressure, ventricular volume the electrocardiogram, and the phonocardiogram

phonocardiogram, which is a recording of the sounds produced by the heart as it pumps It is especially important that the student study in detail the diagram of this figure and understand the causes of all the events illustrated. These are explained as follows·

RELATIONSHIP OF THE ELECTROCARDIOGRAM TO THE CARDIAC CYCLE

The electrocardiogram in Figure 13–5 shows the *P, Q, R, S,* and *T waves,* which will be discussed in Chapters 15, 16, and 17. These are electrical voltages generated by the heart and recorded by the electrocardiograph from the surface of the body. The *P wave* is caused by the *spread of depolarization* through the atria, and this is followed by atrial contraction, which causes a slight rise in the atrial pressure curve immediately after the P wave Approximately 0 16 second after the onset of the P wave, the *QRS waves* appear as a result of depolarization of the ventricles, which initiates contraction of the ventricles and causes the ventricular pressure to begin rising, as also illustrated in the figure. Therefore, the QRS complex begins slightly before the onset of ventricular systole.

Finally, one observes the *ventricular T wave* in the electrocardiogram. This represents the stage of repolarization of the ventricles at which time the ventricular muscle fibers begin to relax. Therefore, the T wave occurs slightly prior to the end of ventricular contraction.

Function of the Atria As Pumps. Blood normally flows continually from the great veins into the atria; approximately 70 per cent of this flow directly through the atria into the ventricles even before the atria contract. Then, atrial contraction usually causes an additional 20 to 30 per cent filling of the ventricles. Therefore, the atria simply function as primer pumps that increase the ventricular pumping effectiveness as much as 30 per cent. Yet, the heart can continue to operate quite satisfactorily under normal resting conditions even without this extra 30 per cent effectiveness because it normally has the capability of pumping 300 to 400 per cent more blood than is required by the body anyway. Therefore, the difference is not likely to be noticed unless a person exercises; then acute signs of heart failure occasionally develop, especially shortness of breath.

Pressure Changes in the Atria — The a, c, and v Waves. In the atrial pressure curve of Figure 13–5 three major pressure elevations called the *a, c, and v atrial pressure waves* can be noted

The *a wave* is caused by atrial contraction. Ordinarily the *right* atrial pressure rises 4 to 6 mm Hg during atrial contraction, whereas the *left* atrial pressure rises about 7 to 8 mm Hg

The *c wave* occurs when the ventricles begin to contract; it is caused partly by slight backflow of blood into the atria at the onset of ventricular contraction, but probably mainly by bulging of the A-V valves toward the atria because of increasing pressure in the ventricles

The *v wave* occurs toward the end of ventricular contraction; it results from slow buildup of blood in the atria while the A-V valves are closed during ventricular contraction. Then, when ventricular contraction is over, the A-V valves open, allowing blood to flow rapidly into the ventricles and causing the v wave to disappear.

FUNCTION OF THE VENTRICLES AS PUMPS

Filling of the Ventricles. During ventricular systole large amounts of blood accumulate in the atria because of the closed A-V valves. Therefore, just as soon as systole is over and the ventricular pressures fall again to their low diastolic values, the high pressures in the atria immediately push the A-V valves open and allow blood to flow rapidly into the ventricles, as shown by the rise of the *ventricular volume curve* in Figure 13–5. This is called the *period of rapid filling of the ventricles*. The atrial pressures fall to within a fraction of a millimeter of the ventricular pressures because the normal A-V valve openings are so large that they offer almost no resistance to blood flow.

The period of rapid filling lasts approximately the first third of diastole. During the middle third of diastole only a small amount of blood normally flows into the ventricles; this is blood that continues to empty into the atria from the veins and passes on through the atria directly into the ventricles. This middle third of diastole, when the inflow of blood into the ventricles is almost at a standstill, is called *diastasis*.

During the latter third of diastole, the atria contract and give an additional thrust to the inflow of blood into the ventricles; this accounts for approximately 20 to 30 per cent of the filling of the ventricles during each heart cycle.

Emptying of the Ventricles During Systole. *Period of Isovolumic (Isometric) Contraction.* Immediately after ventricular contraction begins, the ventricular pressure abruptly rises, as shown in Figure 13–5, causing the A-V valves to close. Then an additional 0.02 to 0.03 second is required for the ventricle to build up sufficient pressure to push the semilunar (aortic and pulmonary) valves open against the pressures in the aorta and pulmonary artery. Therefore, during this period of time, contraction is occurring in the ventricles, but there is no emptying. This period is called the period of *isovolumic* or *isometric contraction*, meaning by these terms that tension is increasing in the muscle but no shortening of the muscle fibers is occurring. (This is not strictly true because there is apex-to-base shortening and circumferential elongation.)

Period of Ejection. When the left ventricular pressure rises slightly above 80 mm Hg (and the right ventricular pressure slightly above 8 mm Hg), the ventricular pressures now push the semilunar valves open. Immediately, blood begins to pour out of the ventricles, with about 70 per cent of the emptying occurring during the first third of the period of ejection and the remaining 30 per cent during the next two thirds. Therefore, the first third is called the *period of rapid ejection* and the last two thirds the *period of slow ejection*.

For a very peculiar reason, the ventricular pressure falls to a value slightly *below* that in the aorta during the period of slow ejection despite the fact that some blood is still leaving the left ventricle. The reason is that the blood flowing out of the ventricle has built up momentum. As this momentum decreases during the latter part of systole, the kinetic energy of the momentum is converted into pressure in the aorta, which makes the arterial pressure slightly greater than the pressure inside the ventricle.

Period of Isovolumic (Isometric) Relaxation. At the end of systole, ventricular relaxation begins suddenly, allowing the intraventricular pressures to fall rapidly. The elevated pressures in the distended large arteries immediately push blood back toward the ventricles, which snaps the aortic and pulmonary valves closed. For another 0.03 to 0.06 second, the ventricular muscle continues to relax, even though the ventricular volume does not change, giving rise to the period of *isovolumic* (or *isometric*) *relaxation*. During this period the intraventricular pressures fall rapidly back to their very low diastolic levels. Then the A-V valves open to begin a new cycle of ventricular pumping.

End-Diastolic Volume, End-Systolic Volume, and Stroke Volume Output. During diastole, filling of the ventricles normally increases the volume of each ventricle to about 120 to 130 ml. This volume is known as the *end-diastolic volume*. Then, as the ventricles empty during systole, the volume decreases about 70 ml, which is called the *stroke volume output*. The remaining volume in each ventricle, about 50 to 60 ml, is called the *end-systolic volume*. The fraction of the end-diastolic volume that is ejected is called the *ejection fraction*—usually equal to about 60 per cent.

When the heart contracts strongly, the end-systolic volume can fall to as little as 10 to 30 ml. On the other hand, when large amounts of blood flow into the ventricles during diastole, their end-diastolic volumes can become as great as 150 to 180 ml in the normal heart. And by both increasing the end-diastolic volume and decreasing the end-systolic volume, the stroke volume output can at times be increased to more than double normal.

FUNCTION OF THE VALVES

The Atrioventricular Valves. The *A-V valves* (the *tricuspid* and the *mitral* valves) prevent backflow of blood from the ventricles to the atria during systole, and the *semilunar valves* (the *aortic* and *pulmonary* valves) prevent backflow from the aorta and pulmonary arteries into the ventricles during diastole. All these valves, which are illustrated in

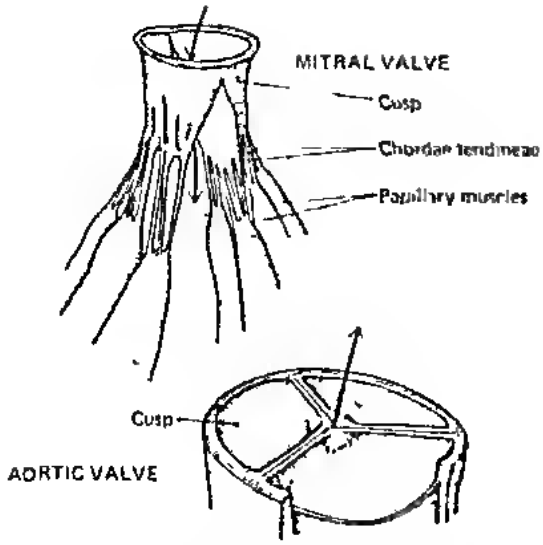


Figure 13–6. Mitral and aortic valves

Figure 13–6, close and open *passively*. That is, they close when a backward pressure gradient pushes blood backward, and they open when a forward pressure gradient forces blood in the forward direction. For obvious anatomical reasons, the thin, filmy A-V valves require almost no backflow to cause closure while the much heavier semilunar valves require rather strong backflow for a few milliseconds.

Function of the Papillary Muscles. Figure 13–6 also illustrates the papillary muscles that attach to the vanes of the A-V valves by the *chordae tendineae*. The papillary muscles contract when the ventricular walls contract, but, contrary to what might be expected, they *do not* help the valves to close. Instead, they pull the vanes of the valves inward toward the ventricles to prevent their bulging too far backward toward the atria during ventricular contraction. If a chorda tendinea becomes ruptured or if one of the papillary muscles becomes paralyzed, the valve bulges far backward, sometimes so far that it leaks severely and results in severe or even lethal cardiac incapacity.

The Aortic and Pulmonary Valves. There are differences between the operation of the aortic and pulmonary valves and that of the A-V valves. First, the high pressures in the arteries at the end of systole cause the semilunar valves to snap to the closed position in comparison with a much softer closure of the A-V valves. Second, the velocity of blood ejection through the aortic and pulmonary valves is far greater than that through the much larger A-V valves. Because of the rapid closure and rapid ejection, the edges of the semilunar valves are subjected to much greater mechanical abrasion than are the A-V valves, which also are supported by the chordae tendineae. It is obvious from the anatomy of the aortic and pulmonary valves, as illustrated in Figure 13–6, that they are well adapted to withstand this extra physical trauma.

THE AORTIC PRESSURE CURVE

The pressure that develops in the aorta depend on many other factors in addition to contraction the heart. The pulsatile aortic pressure curve w be discussed in detail in Chapter 19, but it desirable now simply to point out those features the curve that relate especially to the cardiac cycl

When the left ventricle contracts, the ventric lar pressure rises rapidly until the aortic val opens. Then the pressure in the ventricle ris much less thereafter, as illustrated in Figu 13–5, because blood immediately flows out of t ventricle into the aorta.

The entry of blood into the arteries causes t walls of these arteries to stretch and the pressu to rise. Then, at the end of systole, after the le ventricle stops ejecting blood and the aortic val closes, the elastic recoil of the arteries maintai a high pressure in the arteries even during dia tole.

A so-called *incisura* occurs in the aortic pressu curve when the aortic valve closes. This is cause by a short period of backward flow of blood imm diately prior to closure of the valve, followed th by sudden cessation of the backflow.

After the aortic valve has closed, pressure in t aorta falls slowly throughout diastole becau blood stored in the distended elastic arteries flo continually through the peripheral vessels back the veins. Before the ventricle contracts again, t aortic pressure usually falls to approximately mm Hg (diastolic pressure), which is two thir the maximal pressure of 120 mm Hg (systol pressure) occurring in the aorta during ventricul contraction.

The pressure curve in the pulmonary artery similar to that in the aorta, except that the pre sures are only about one sixth as great, as will discussed in Chapter 24.

RELATIONSHIP OF THE HEART SOUNDS TO HEART PUMPING

When listening to the heart with a stethoscope, o does not hear the opening of the valves, for this is relatively slowly developing process that makes noise. However, when the valves close, the vanes of t valves and the surrounding fluids vibrate under t influence of the sudden pressure differentials that d velop, giving off sound that travels in all directio through the chest.

When the ventricles first contract, one hears a sou that is caused by closure of the A-V valves. The vibrati is low in pitch and relatively long continued and known as the *first heart sound*. When the aortic a pulmonary valves close, one hears a relatively rap snap, for these valves close extremely rapidly, and surroundings vibrate for only a short period of tim This sound is known as the *second heart sound*. precise causes of these sounds will be discussed Chapter 27 in relation to auscultation.

Occasionally, one can hear an *atrial sound* when atria beat, because of vibrations associated with the flo

blood into the ventricles. Also, a *third heart sound* metimes occurs at about the end of the first third of astole. This is believed to be caused by blood flowing a rumbling motion into the almost-filled ventricles he atrial sound and the third heart sound can usually e recorded with special recording instruments but can e heard with the stethoscope only with great difficulty The Terms "Systole" and Diastole" in Clinical sage. Strictly speaking, systole means "contraction." herefore, physiologically it is probably best that systole e considered to begin approximately with the closure f the A-V valves and to end approximately with the pening of the A-V valves.

Clinically, it is not possible to determine when the A- valves open. Yet, because only a short interval elapses etween the closure of the aortic and pulmonary valves nd the opening of the A-V valves, the clinician meas- res systole as the time interval between the first heart ound and the second heart sound or, in other words, he time interval between closure of the A-V valves and losure of the semilunar valves. Diastole is considered o be the interval between closure of the aortic and ulmonary valves and closure of the A-V valves

VORK OUTPUT OF THE HEART

The work output of the heart is the amount of energy hat the heart converts to work while pumping blood nto the arteries This is in two forms. First, by far the najor proportion is used to move the blood from the low ressure veins to the high pressure arteries This is *otential energy of pressure.* Second, a minor proportion f the energy is used to accelerate the blood to its velocity ejection through the aortic and pulmonary valves This is *kinetic energy of blood flow*

Stroke Work Output. The work performed by the eft ventricle to raise the pressure of the blood during ach heartbeat is called the left ventricular stroke work utput and is equal to *stroke volume output × (left entricular mean ejection pressure* minus *left atrial pres- sure).* Likewise, the work performed by the right ven- ricle to raise the pressure of the blood is called the ight ventricular stroke work output. When pressure is expressed in *dynes per square centimeter* and stroke olume output in *milliliters,* the work output is in *ergs* Right ventricular work output is usually about one sixth the work output of the left ventricle because of the difference in systolic pressure against which the two ventricles must pump.

The Kinetic Energy of Blood Flow. The work out- put of each ventricle required to create kinetic energy of blood flow is proportional to the mass of blood ejected times the square of velocity of ejection. That is,

$$\text{Kinetic energy} = \frac{mv^2}{2}$$

When the mass is expressed in *grams* of blood ejected and the velocity in *centimeters per second,* the work output is in *ergs.*

Ordinarily, the work output of the left ventricle re- quired to create kinetic energy of blood flow is only about 1 per cent of the total work output of the ventricle and therefore is ignored in the calculation of the total stroke work output. Most of this energy is required to cause the rapid acceleration of blood during the first quarter of systole. In certain abnormal conditions, such as aortic stenosis, in which the blood flows with great velocity through the stenosed valve, as much as 50 per cent of the total work output may be required to create kinetic energy of blood flow.

ENERGY FOR CARDIAC CONTRACTION

Heart muscle, like skeletal muscle, utilizes chemical energy to provide the work of contraction. This energy is derived mainly from metabolism of fatty acids and to a lesser extent from metabolism of other nutrients, especially lactate and glucose. The different reactions that liberate this energy will be discussed in Chapters 67 and 68

The amount of energy expended by the heart is determined by three important factors: (1) the tension in the heart muscle at each instant during contraction, (2) the fraction of the time that the tension is maintained, and (3) the rate of muscle shortening at each instant while the tension is maintained The first two factors, the tension and the time, are much more important than the rate of shortening Therefore, it is frequently stated that the energy expended is approximately proportional to *tension* times *time*

Efficiency of Cardiac Contraction. During muscular contraction most of the chemical energy is converted into heat and a small portion into work output. The ratio of work output to chemical energy expenditure is called the efficiency of cardiac contraction, or simply *efficiency of the heart* The maximum efficiency of the normal heart is between 20 and 25 per cent

REGULATION OF CARDIAC FUNCTION

When a person is at rest, the heart pumps only 4 to 6 liters of blood each minute. However, during severe exercise it may be required to pump as much as four to seven times this amount. The present section discusses the means by which the heart can adapt to such extreme increases in cardiac output.

The two basic means by which the volume pumped by the heart is regulated are (1) intrinsic regulation of pumping in response to changes in volume of blood flowing into the heart, and (2) reflex control of the heart by the autonomic nervous system.

INTRINSIC REGULATION OF CARDIAC PUMPING—THE FRANK-STARLING LAW OF THE HEART

In Chapter 23 we shall see that one of the major factors determining the amount of blood pumped by the heart each minute is the rate of blood flow into the heart from the veins, which is called *venous return.* That is, each peripheral tissue of the body controls its own blood flow, and whatever amount of blood flows through all the peripheral tissues returns by way of the veins to the right atrium. The heart in turn automatically pumps this incoming blood on into the systemic arteries

so that it can flow around the circuit again. Thus, the heart must adapt from moment to moment or even second to second to widely varying inputs of blood, sometimes falling as low as 2 to 3 liters per minute and at other times rising to as high as 25 or more liters per minute.

This intrinsic ability of the heart to adapt to changing loads of inflowing blood is called the *Frank-Starling law of the heart*, in honor of Frank and Starling, two great physiologists of nearly a century ago. Basically, the Frank-Starling law states that the greater the heart is filled during diastole, the greater will be the quantity of blood pumped into the aorta. Or another way to express this law is: *Within physiological limits, the heart pumps all the blood that comes to it without allowing excessive damming of blood in the veins*

Mechanism of the Frank-Starling Law. The principal mechanism by which the heart adapts to changing inflow of blood is the following. When the cardiac muscle becomes stretched an extra amount, as it does when extra amounts of blood enter the heart chambers, the stretched muscle contracts with a greatly increased force, thereby automatically pumping the extra blood into the arteries This ability of stretched muscle to contract with increased force is characteristic of all striated muscle, not simply of cardiac muscle Referring back to Chapter 11, one will see that stretching a skeletal muscle, within its physiological limit, also increases its force of contraction. This increased force of contraction is probably caused by the fact that the actin and myosin filaments are brought to a more nearly optimal degree of interdigitation for achieving contraction.

Effect of Heart Rate. In addition to the important effect of stretching the heart muscle, still another factor increases heart pumping when its volume is increased. Stretch of the right atrial wall directly increases the heart rate by as much as 10 to 30 per cent; this too helps increase the amount of blood pumped each minute, though its contribution is much less than that of the Frank-Starling mechanism

Lack of Effect on Cardiac Output of Changes in Arterial Pressure Load. One of the most important consequences of the Frank-Starling law of the heart is that, within reasonable limits, changes in arterial pressure load against which the heart pumps have almost no effect on the rate at which blood is pumped by the heart each minute (the cardiac output) This effect is illustrated in Figure 13–7, which is a curve extrapolated to the human being from data in dogs in which the arterial pressure was progressively changed by constricting the arteries while the cardiac output was measured simultaneously. The significance of this effect is the following. Regardless of the arterial pressure, the most important factor determining the amount of blood pumped by the heart is still the rate of entry of blood into the heart.

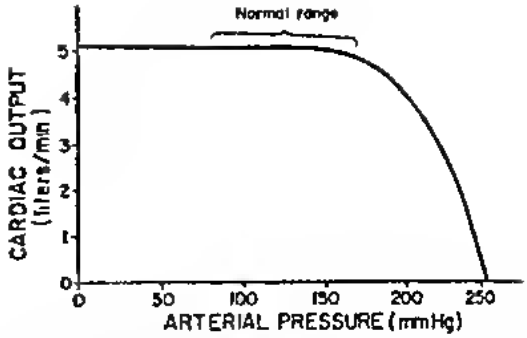


Figure 13–7. Constancy of cardiac output even in the fac wide changes in arterial pressure Only when the arterial pres rises above the normal operating pressure range does the p sure load cause the heart to begin to fail.

Figure 13–7 shows that when the arterial p sure rises above approximately 170 mm Hg, arterial pressure load then does cause the hear begin to fail. However, this figure also shows t the normal daily range of arterial pressure between approximately 80 and 170 mm Hg, ag emphasizing that within the normal operat range, the output of the heart is independen changes in arterial pressure.

The independence of the cardiac output changes in arterial pressure load is partly cau by the fact that the heart is a two-stage pump, the following reasons: Even if tha left ventr were to begin to fail moderately, this would r the left atrial pressure a few millimeters Hg. T elevation of left atrial pressure then would r the pulmonary arterial pressure only a fractio a millimeter because left atrial pressure is transmitted backward through the lungs to a gr extent until this pressure becomes abnorm elevated. Therefore, the right ventricle experier almost no change in load and continues to pu normal quantities of blood into the lungs, forc the left ventricle (even though it might be be ning to fail) to continue to pump a normal car output.

Relationship of the Ventricular Volume to V tricular Pressure. One can learn much about functional capability of the heart by studying the r tionship between ventricular volume and ventric pressure. The two solid curves of Figure 13–8 illust this relationship during both diastole and systole for left ventricle.

The diastolic pressure curve is determined by fil the heart with progressively greater quantities of b and then measuring the diastolic pressure immedia before ventricular contraction occurs, which is the *diastolic pressure* of the ventricle.

The systolic pressure curve is determined by prev ing any outflow of blood from the heart and measu the maximum systolic pressure that is achieved du ventricular contraction at each volume of filling

It is very clear that until the volume of the ve rises above about [illegible] c press increase greatly. Above this volume, the diastolic sure does increase rapidly [illegible] tissue in the heart that will stretch no more, and parti because the pericardium that surrounds the heart

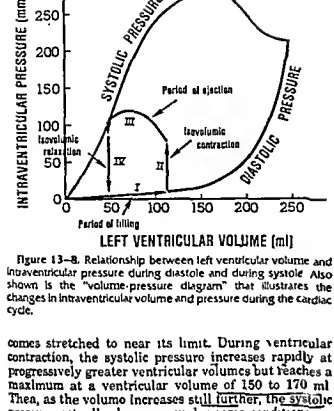


Figure 13–8. Relationship between left ventricular volume and intraventricular pressure during diastole and during systole. Also shown is the "volume-pressure diagram" that illustrates the changes in intraventricular volume and pressure during the cardiac cycle.

comes stretched to near its limit. During ventricular contraction, the systolic pressure increases rapidly at progressively greater ventricular volumes but reaches a maximum at a ventricular volume of 150 to 170 ml. Then, as the volume increases still further, the systolic pressure actually decreases under some conditions, as illustrated by the falling systolic pressure curve, because at these very great volumes the actin and myosin filaments of the cardiac muscle fibers are actually pulled apart enough for the strength of cardiac fiber contraction to become less than optimal.

Note especially in the figure that the maximum systolic pressure for the normal unstimulated left ventricle is between 250 and 300 mm Hg, but this varies widely with strength of the heart. For the normal right ventricle it is between 60 and 80 mm Hg.

The "Volume-Pressure Diagram" During the Cardiac Cycle, and Cardiac Work Output. The colored curves in Figure 13–8 form a loop called the *volume-pressure diagram* of the cardiac cycle for the left ventricle. It is divided into four separate phases.

Phase I: *Period of filling.* This phase in the volume-pressure diagram begins at a ventricular volume of about 45 ml and a diastolic pressure very near 0 mm Hg. Forty-five ml is the amount of blood that remains in the ventricle after the previous heartbeat, and is called the *end-systolic volume.* As new blood enters the ventricle from the atrium, the volume increases normally to about 115 ml, called the *end-diastolic volume,* an increase of 70 ml. Therefore, the volume-pressure diagram during phase I extends along the curve labeled "I," with the volume increasing to 115 ml and the diastolic pressure rising to about 5 mm Hg.

Phase II. *Period of isovolumic contraction.* During isovolumic contraction, the volume of the ventricle does not change. However, the pressure inside the ventricle rises to equal the pressure in the aorta at the end of diastole, a pressure value of about 80 mm Hg, as depicted by the curve labeled "II."

Phase III: *Period of ejection.* During ejection, the systolic pressure rises even higher because of still more contraction of the heart. On the other hand, the volume of the ventricle decreases. Therefore, the curve labeled "III" traces the changes in volume and systolic pressure during this period of ejection.

Phase IV: *Period of isovolumic relaxation.* At the end of the period of ejection, the semilunar valves of the ventricles close, and the ventricular pressure falls back to the diastolic pressure level. The curve labeled "IV" traces this decrease in intraventricular pressure without any change in volume. Thus, the ventricle returns to its starting point, with about 45 ml of blood left in the ventricle and at an atrial pressure usually very close to 0 mm Hg.

Students well trained in the basic principles of physics will recognize that the area subtended by this volume-pressure diagram is equal to the *net work output* of the ventricle during its contraction cycle. In experimental studies of cardiac contraction, this diagram is therefore used *for calculating cardiac work output.*

When the heart pumps large quantities of blood, the work diagram becomes much larger in area. It extends far to the right because the ventricle fills with more blood during diastole; it rises much higher because the ventricle contracts with greater pressure; and it usually extends farther to the left because the ventricle contracts to a smaller volume—especially if the ventricle is stimulated to increased activity by the sympathetic nervous system.

The Concepts of "Preload" and "Afterload." In assessing the contractile properties of muscle, it is important to specify the degree of stretch of the muscle when it begins to contract, which is called the *preload,* and also to specify the load against which the muscle exerts its contractile force, which is called the *afterload.*

For cardiac contraction, the preload is considered to be the volume of blood in the ventricle at the end of diastole, that is, the end-diastolic volume. However, sometimes this preload is expressed as the end-diastolic pressure that fills the ventricle.

The afterload of the ventricle is the pressure in the arteries leading from the ventricles. In Figure 13–8 this corresponds to the changing systolic pressure described by the Phase III curve of the volume-pressure diagram.

The importance of the concepts of preload and afterload is that in many abnormal functional states of the heart or circulation the degree of filling of the ventricles (the preload) and/or the arterial pressures against which the ventricles must contract (the afterload) are severely altered from the normal.

Ventricular Function Curves

One of the best ways to express the functional ability of the ventricles to pump blood is by ventricular function curves, as shown in Figures 13–9 and 13–10. Figure 13–9 illustrates a type of ventricular function curve called the *stroke work output curve.* Note that as the atrial pressures increase, the stroke work output also increases until it reaches the limit of the heart's ability.

Figure 13–10 illustrates another type of ventricular function curve called the *minute ventricular output curve.* These two curves represent function of the two ventricles of the human heart based on data extrapolated from lower [illegible]

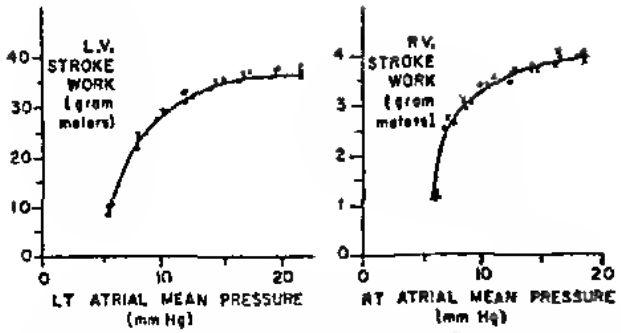


Figure 13–9. Left and right ventricular function curves in a dog, depicting ventricular stroke work output as a function of left and right mean atrial pressures (Curves reconstructed from data in Sarnoff *Physiol. Rev.*, 35 107, 1955)

atrial pressure rises, the respective ventricular volume output per minute also increases.

Thus, ventricular function curves are another way of expressing the Frank-Starling law of the heart. That is, as the ventricles fill to higher atrial pressures, the strength of cardiac contraction increases, causing the heart to pump increased quantities of blood into the arteries. In later chapters we shall see that ventricular output curves are exceedingly important in analyzing overall function of the circulation, for it is by such means that one can express the quantitative capabilities of the heart as a pump.

CONTROL OF THE HEART BY PARASYMPATHETIC AND SYMPATHETIC NERVES

The heart is well supplied with both sympathetic and parasympathetic (vagal) nerves, as illustrated in Figure 13–11. These nerves affect cardiac pumping in two ways: (1) by changing the heart rate, and (2) by changing the strength of contraction of the heart. The effect of nerve stimulation on heart rate and rhythm will be discussed in the following chapter. For the present, suffice it to say that parasympathetic stimulation decreases heart rate and sympathetic stimulation increases heart rate. The range of control is from as little as 20 to 30 heart beats per minute with maximum vagal stim-

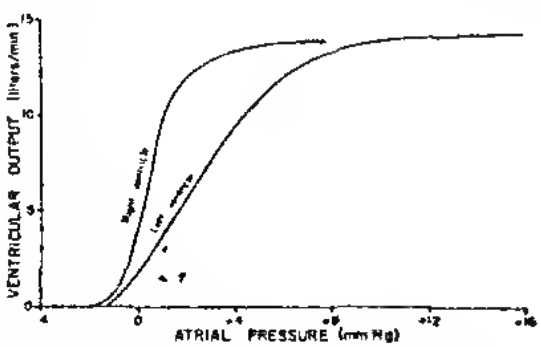


Figure 13–10. Approximate normal right and left ventricular output curves for the human heart as extrapolated from data obtained in dogs

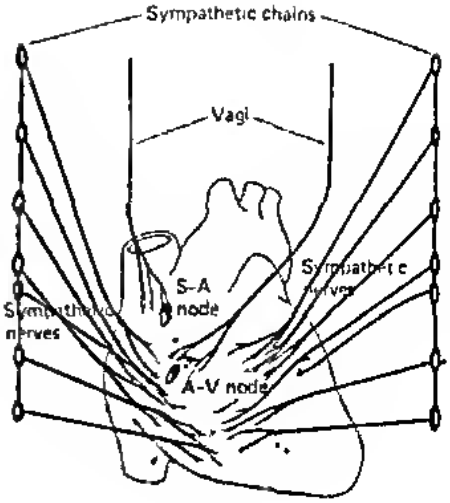


Figure 13–11. The cardiac nerves

ulation to as high as 250 beats per minute v maximum sympathetic stimulation.

Effect of Heart Rate on Function of Heart as a Pump. In general, the more times heart beats per minute, the more blood it pump, but there are important limitations to t effect. For instance, once the heart rate rises ab a critical level, the heart strength itself decrea presumably because of overutilization of metab substrates in the cardiac muscle. In addition, period of diastole between the contractions comes so reduced that blood does not have tim flow adequately from the atria into the ventric For these reasons, when the heart rate is increa artificially by electrical stimulation, the heart its peak ability to pump large quantities of bl at a heart rate between 100 and 150 beats minute. On the other hand, when its rate is creased by sympathetic stimulation, it reaches peak ability to pump blood at a heart rate betw 170 and 250 beats per minute. The reason for t difference is that sympathetic stimulation not o increases the heart rate but also increases he strength as well, while at the same time decreas the duration of systolic contraction and allow more time for filling during diastole.

Nervous Regulation of Contractile Stren of the Heart. The two atria are especially v supplied with large numbers of both sympath and parasympathetic nerves, but the ventri are supplied mainly by sympathetic nerves and fewer parasympathetic fibers. In general, sym thetic stimulation increases the strength of he muscle contraction, whereas parasympath stimulation decreases the strength of contracti

Under normal conditions the sympathetic ne fibers to the heart continually discharge at a s rate that maintains a strength of ventricular c tract about 20 per cent above its strength wi no sympathetic stimulation at all. Therefore, o method by which the nervous system can decrea the strength of ventricular contraction is simpl

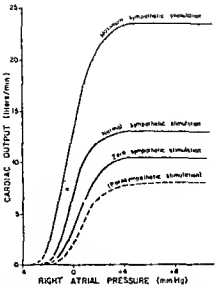


Figure 13–12. Effect on the cardiac output curve of different egrees of sympathetic and parasympathetic stimulation

) slow or stop the transmission of sympathetic mpulses to the heart. On the other hand, maximal ympathetic stimulation can increase the strength f ventricular contraction to approximately 100 er cent greater than normal.

Maximal parasympathetic stimulation of the eart decreases ventricular contractile strength bout 30 per cent. Thus, the parasympathetic effect s, by contrast with the sympathetic effect, relatively small. (However, parasympathetic stimulation can greatly decrease the heart rate, as discussed in the following chapter.)

Effect of Sympathetic or Parasympathetic Stimulation on the Cardiac Function Curve. Figure 13–12 illustrates four separate cardiac function curves. These are much the same as the ventricular function curves of Figures 13–9 and 13–10, except that they represent function of the entire heart rather than of a single ventricle; they show the relationship between the right atrial pressure at the input of the heart from the great veins and cardiac output from the left ventricle into the aorta.

The curves of Figure 13–12 demonstrate that at any given right atrial pressure, the cardiac output increases with increasing sympathetic stimulation and decreases with increasing parasympathetic stimulation. It should be remembered that the changes in output caused by nerve stimulation are brought about by both *changes in heart rate* and *changes in contractile strength of the heart*, for both of these affect cardiac output.

EFFECT OF HEART DEBILITY ON CARDIAC FUNCTION—THE HYPOEFFECTIVE HEART

Any factor that damages the heart, whether it be damage to the myocardium, to the valves, to the conducting system, or otherwise, is likely to make the heart a poorer pump, and the heart under these conditions is called a *hypoeffective heart*. Figure 13–13 illustrates by the very dark curve the normal cardiac function curve and by the three curves below this the effect of different degrees of hypoeffectiveness on cardiac function. Obviously, the more serious the damage, the less will be the cardiac output at any given right atrial pressure.

Different factors that can cause a hypoeffective heart include:

- Myocardial infarction
- Valvular heart disease
- Vagal stimulation of the heart
- Inhibition of the sympathetics to the heart
- Congenital heart disease
- Myocarditis
- Cardiac anoxia
- Diphtheritic or other types of myocardial damage

EFFECT OF CHRONIC INCREASE IN HEART WORK LOAD—THE HYPEREFFECTIVE HEART

Chronic increase in the work load of the heart over a period of many weeks or months leads to hypertrophy of the cardiac muscle and also to enlargement of the ventricular chambers. As a result, the overall effectiveness of the heart as a pump increases. The upper three function curves of Figure 13–13 illustrate the effect of different degrees of cardiac hypereffectiveness on the cardiac function curve; maximal degrees of hypereffectiveness can increase pumping by the heart more than 100 per cent.

Other factors besides hypertrophy that can cause a hypereffective heart are

- Sympathetic stimulation (50 to 100 per cent increase);
- Inhibition of parasympathetic stimulation (10 to 20 per cent increase)

EFFECT OF VARIOUS IONS ON HEART FUNCTION

In the discussion of membrane potentials in Chapter 10, it was pointed out that three particular cations—

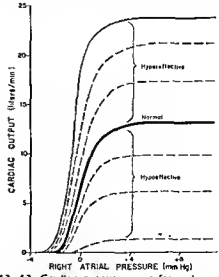


Figure 13–13. Cardiac output curves for various degrees of hypo- and hypereffective hearts. (from Guyton, Jones, and Coleman Circulatory Physiology, Cardiac Output and Its Regulation Philadelphia, W. B. Saunders Company, 1973.)

potassium, sodium, and calcium—all have marked effects on membrane potentials and action potentials, and in Chapter 11 it was noted that calcium ions play an especially important role in initiating the muscle contractile process. Therefore, it is to be expected that the concentrations of these three ions in the extracellular fluids will also have important effects on cardiac function.

The action of individual ions on the heart can be studied by perfusing the isolated heart, as illustrated in Figure 13–14, using the so-called *Langendorf preparation*. In making this preparation, the heart is removed rapidly from the animal's body, and the aorta is cannulated immediately. The cannula is connected to a perfusion system so that the fluid in the aorta will be under pressure. The pressure causes the fluid to flow through the coronary vessels. When the perfusion fluid contains appropriate nutrients (glucose and oxygen) and ions, and the heart is kept appropriately warmed, the perfused heart will beat for many hours.

Effect of Potassium Ions. Excess potassium in the extracellular fluids causes the heart to become extremely dilated and flaccid and slows the heart rate. Very large quantities can also block conduction of the cardiac impulse from the atria to the ventricles through the A-V bundle. Elevation of potassium concentration to only 8 to 12 mEq/liter—two to three times the normal value—can cause such weakness of the heart and abnormal rhythm that this can cause death.

All these effects are caused by the fact that a high potassium concentration in the extracellular fluids causes a decreased resting membrane potential in the cardiac muscle fibers, as explained in Chapter 10. And, as the membrane potential decreases, the intensity of the action potential also decreases, which makes the contraction of the heart progressively weaker, for the strength of the action potential determines to a great extent the strength of contraction.

Effect of Calcium Ions. An excess of calcium ions causes effects almost exactly opposite of those of potassium ions, causing the heart to go into spastic contraction. This is caused by the direct effect of calcium io in exciting the cardiac contractile process, as explai earlier in the chapter. Conversely, a deficiency of cium ions causes cardiac flaccidity, similar to the ef of potassium.

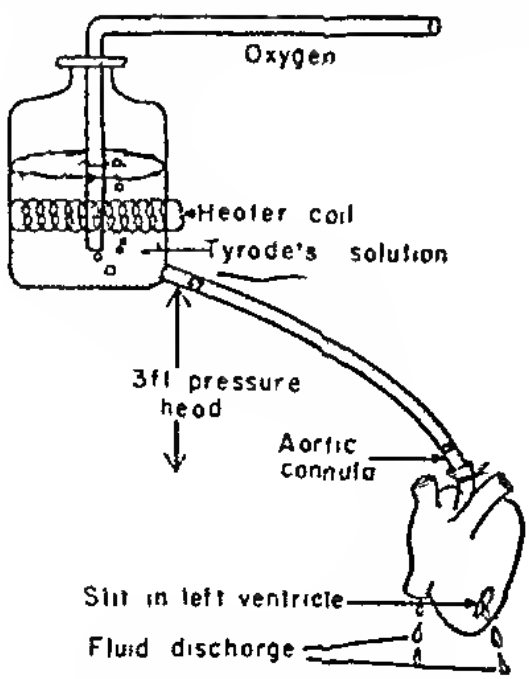


Figure 13–14. System for perfusing the heart

Effect of Sodium Ions. An excess of sodium i depresses cardiac function, an effect similar to tha potassium ions but for an entirely different rea Sodium ions compete with calcium ions at some unexplained point in the excitatory or the contrac process of heart muscle in such a way that the gre the sodium ion concentration in the extracellular flu the less the effectiveness of the calcium ions in cau contraction when an action potential occurs.

However, from a practical point of view, the sodi ion concentration in the extracellular fluids prob never becomes high enough, even in serious patholog conditions, to cause significant change in car strength because of the extreme degree of effectiv of the sodium concentration–regulating mechanism the other hand, very low sodium concentration, as oc in water intoxication, often causes death becaus cardiac fibrillation, a phenomenon explained in following chapter.

EFFECT OF TEMPERATURE ON THE HEART

The effect of temperature on the heart can also studied using the perfusion system of Figure 13- Increased temperature causes greatly increased h rate, and decreased temperature causes greatly creased rate. These effects presumably result from heat causing increased permeability of the muscle m brane to the ions, resulting in acceleration of the excitation process.

Contractile strength of the heart is often enhan temporarily by a moderate increase in temperature, prolonged elevation of the temperature exhausts metabolic systems of the heart and causes weakness

THE HEART-LUNG PREPARATION

A method frequently used to demonstrate the c bilities of the heart as a pump is the heart-lung pr ration illustrated in Figure 13–15. In this preparat blood leaving the heart by way of the aorta is dive into an external system of tubes and then back ag into the right atrium. The only parts of the anim body that remain alive are the heart and lungs blood is oxygenated by artificial respiration of the lu and glucose is added periodically to the blood to su nutrition for cardiac contraction.

As blood flows through the external circuit, the ri atrial pressure, aortic pressure, and blood flow are me ured. Also, an adjustable venous reservoir allows pressure in the right atrium to be increased or decrea A screw clamp is provided on the main outflow t from the heart so that the arterial resistance ca increased or decreased. Finally, a special reservoir c taining air in its upper part and blood in its lower p is used as an "elastic buffer" *(windkessel)* so blood pressure will not rise and fall excessively each contraction of the heart. Some of the cardiac function that can be demonstrated with heart-lung preparation are the following:

First, the *effect of end-diastolic volume or atrial pre sure on stroke volume output* can be demonstrated

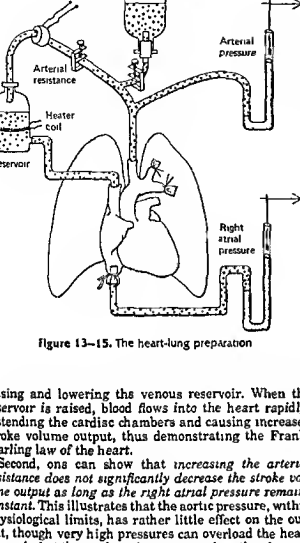


Figure 13–15. The heart-lung preparation

raising and lowering the venous reservoir. When the reservoir is raised, blood flows into the heart rapidly, distending the cardiac chambers and causing increased stroke volume output, thus demonstrating the Frank-Starling law of the heart.

Second, one can show that *increasing the arterial resistance does not significantly decrease the stroke volume output as long as the right atrial pressure remains constant.* This illustrates that the aortic pressure, within physiological limits, has rather little effect on the output, though very high pressures can overload the heart so much that the stroke volume output does then become reduced.

Third, the *effect of different factors on heart rate* can be studied, such as a slight increase in heart rate that occurs when increased inflow of blood to the heart stretches the right atrial wall, an increase in heart rate with increasing temperature, an increase in heart rate when epinephrine is injected into blood perfusing the heart, a decrease in heart rate when acetylcholine is injected, and changes in heart rate when the nerves to the heart are stimulated.

ASSESSMENT OF CONTRACTILITY

Though it is very easy to determine the heart rate by simply timing the pulse, it has always been *difficult* to determine the strength of contraction of the heart, commonly called *cardiac contractility.* Very often the change in contractility is exactly opposite to the change in heart rate. Indeed, this effect occurs almost invariably in heart-debilitating diseases.

One of the ways in which cardiac contractility can be determined with great precision is to record one or more of the cardiac function curves. However, this can be done only in experimental animals, and even then only with considerable difficulty. Therefore, many physiologists and clinicians have searched for methods to assess the cardiac contractility in a simple way. One of these methods is to determine the so-called dP/dt.

dP/dt as a Measure of Cardiac Contractility. dP/dt means the *rate of change of the ventricular pressure* with respect to time. The dP/dt record is generated by an electronic computer that *differentiates* the ventricular pressure wave, thus giving a record of the rate of change of the ventricular pressure. Figure 13–16 illustrates two separate recordings of the ventricular pressure wave as well as simultaneous recordings (in color) of the dP/dt. In the upper part of the figure the heart was beating normally, and in the lower part the heart had been stimulated by isoproterenol, a drug that has essentially the same effect on the heart as sympathetic stimulation, as will be discussed in more detail in Chapter 57.

Note in the upper record that at the same time that the ventricular pressure is increasing at its most rapid rate, the recording of the dP/dt record also reaches its greatest height. On the other hand, at the time that the ventricular pressure is falling most rapidly, the dP/dt record reaches its lowest level. When the ventricular pressure is neither rising nor falling, the dP/dt record is at zero value.

Experimental studies have shown that the rate of rise of ventricular pressure, the dP/dt, in general correlates very well with the strength of contraction of the ventricle. This effect is illustrated by a comparison of the dP/dt record in the upper part of Figure 13–16 with that in the lower part. Note that the peak value for dP/dt in the upper record is only 1800 mm Hg per second, whereas in the lower record it rises to approximately 2600 mm Hg per second, illustrating the stimulatory effect of isoproterenol on the contractility of the ventricle. Thus, the *peak* dP/dt is often used as a means for comparing the contractilities of hearts in different functional states.

Unfortunately, the quantitative value for peak dP/dt is also affected by other factors that are not related to cardiac contractility. For instance, the value is increased by both increased input pressure to the left ventricle (the end-diastolic ventricular pressure), which is the *preload* of the ventricle, and the pressure in the aorta

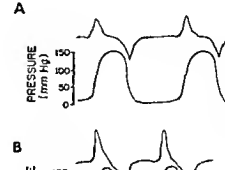


Figure 13–16. Simultaneous recordings of ventricular pressure and dP/dt. *A* shows results from a normal heart and *B*, from a heart stimulated by isoproterenol. (Modified from Mason et al., in Sodeman and Sodeman (eds.) Pathologic Physiology, 6th ed. Philadelphia, W B Saunders Company, 1979.)

against which the heart is pumping the blood, called the *afterload*. Therefore, it is often difficult to use dP/dt as a measure of contractility in comparing hearts from one person to another because one of these factors may differ. For this reason other quantitative measures have also been used in attempts to assess cardiac contractility. One of these has been to use dP/dt divided by the instantaneous pressure in the ventricle, or (dP/dt)/P. Another has been to use V_{max}, which is the maximum rate of shortening of the heart muscle at zero intraventricular pressure, a value almost impossible to measure directly but that can be estimated from other measurements of heart function.

REFERENCES

Alpert, N. R., *et al.* Heart muscle mechanics. *Annu. Rev. Physiol.*, *41*:521, 1979.

Baan, J., *et al.* (eds.) Cardiac Dynamics. Boston: M. Nijhoff, 1979.

Bishop, V. S., *et al.* Factors influencing performance. *Int. Rev. Physiol.*, *9*:239, 1976.

Brady, A. J. Mechanical properties of cardiac fibers. *In* Berne, R. M., *et al.* (eds.): Handbook of Physiology, Sec. 2, Vol. 1. Baltimore, Williams & Wilkins, 1979, p. 533.

Braunwald, E., and Ross, J., Jr. Control of cardiac performance. *In* Berne, R. M., *et al.* (eds.): Handbook of Physiology, Sec. 2, Vol. 1. Baltimore, Williams & Wilkins, 1979, p. 533.

Cowley, A. W., Jr., and Guyton, A. C.: Heart rate as a determinant of cardiac output in dogs with arteriovenous fistula. *Am. J. Cardiol.*, *28*:321, 1971.

Ellis, D.: Na/Ca exchange in cardiac tissue. *Adv. Myocardiol.*, *5*:295, 1985.

Fabiato, A., and Fabiato, F. Calcium and cardiac excitation-contraction coupling. *Annu. Rev. Physiol.*, *41*:473, 1979.

Gadsby, D. C. The Na/K pump of cardiac cells. *Annu. Rev. Biophys. Bioeng.*, *13*:373, 1984.

Gevers, W. Protein metabolism of the heart. *J. Mol. Cell. Cardiol.*, *16*:3, 1984.

Gilmour, R. F., Jr., and Zipes, D. P. Slow inward current and cardiac arrhythmias. *Am. J. Cardiol.*, *55*:89B, 1985.

Guyton, A. C. Determination of cardiac output by equating venous return curves with cardiac response curves. *Physiol. Rev.*, *35*:123, 1955.

Guyton, A. C., *et al.* Circulatory Physiology: Cardiac Output and Its Regulation, 2nd Ed. Philadelphia, W. B. Saunders Co., 1973.

Hoffman, B. B., and Lefkowitz, R. J. Adrenergic receptors in the heart. *Annu. Rev. Physiol.*, *44*:475, 1982.

Jewell, B. R.: The physiology of cardiac muscle contraction. *In* Dickinson, C. J., and Marks, J. (eds.): Developments in Cardiovascular Medicine. Lancaster, England, MTP Press, 1978, p. 129.

Johnson, E. A.: Force-interval relationship of cardiac muscle. *In* Berne, R. M., *et al.* (eds.): Handbook of Physiology, Sec. 2, Vol. 1. Baltimore, Williams & Wilkins, 1979, p. 475.

Langer, G. A., *et al.* The myocardium. *Int. Rev. Physiol.*, *9*:191, 1976.

Lüllmann, H., and Peters, T.: Action of Cardiac Glycosides on the Excitation-Contraction Coupling in Heart Muscle. New York, Fischer, 1979.

Matsubara, I.: X-ray diffraction studies of the heart. *Annu. Rev. Biophys. Bioeng.*, *9*:81, 1980.

Mirsky, I.: Elastic properties of the myocardium: A quantitative approach with physiological and clinical applications. *In* Berne, R. M., *et al.* (eds.): Handbook of Physiology, Sec. 2, Vol. 1. Baltimore, Williams & Wilkins, 1979, p. 497.

Navaratnam, V.: The structure of cardiac muscle. *In* Dickinson, C. J., and Marks, J. (eds.): Developments in Cardiovascular Medicine. Lancaster, England, MTP Press, 1978, p. 119.

Page, E., and Shibata, Y.: Permeable junctions between cardiac cells. *Annu. Rev. Physiol.*, *43*:431, 1981.

Parmley, W. W., and Talbot, W.: Heart as a pump. *In* Berne, R. M., *et al.* (eds.): Handbook of Physiology, Sec. 2, Vol. 1. Baltimore, Williams & Wilkins, 1979, p. 429.

Sanguinetti, M. C., and Kass, R. S.: Regulation of cardiac calcium channel current and contractile activity by the dihydropyridine BAY K 8644 is voltage-dependent. *J. Mol. Cell. Cardiol.*, *16*:667, 1984.

Sarnoff, S. J.: Myocardial contractility as described by ventricular function curves. *Physiol. Rev.*, *35*:107, 1955.

Sommer, J. R., and Johnson, E. A.: Ultrastructure of cardiac muscle. *In* Berne, R. M., *et al.* (eds.): Handbook of Physiology, Sec. 2, Vol. 1. Baltimore, Williams & Wilkins, 1979, p. 113.

Starling, E. H.: The Linacre Lecture on the Law of the Heart. London: Longmans Green & Co., 1918.

Sugden, P. H.: The effects of hormonal factors on cardiac protein turnover. *Adv. Myocardiol.*, *5*:105, 1985.

Sugimoto, T., *et al.*: Quantitative effect of low coronary pressure on left ventricular performance. *Jap. Heart J.*, *9*:46, 1968.

Sugimoto, T., *et al.*: Effect of maximal work load on cardiac function. *Jap. Heart J.*, *14*:146, 1973.

Tada, M., and Katz, A. M.: Phosphorylation of the sarcoplasmic reticulum and sarcolemma. *Annu. Rev. Physiol.*, *44*:401, 1982.

Vary, T. C., *et al.*: Control of energy metabolism of heart muscle. *Annu. Rev. Physiol.*, *43*:419, 1981.

Weisfeldt, M. L. (ed.): The Heart in Old Age: Its Function and Response to Stress. New York, Raven Press, 1980.

Winegrad, S.: Electromechanical coupling in heart muscle. *In* Berne, R. M., *et al.* (eds.): Handbook of Physiology, Sec. 2, Vol. 1. Baltimore, Williams & Wilkins, 1979, p. 293.

Winegrad, S.: Calcium release from cardiac sarcoplasmic reticulum. *Annu. Rev. Physiol.*, *41*:451, 1982.

Winegrad, S.: Regulation of cardiac contractile proteins: Correlation between physiology and biochemistry. *Circ. Res.*, *55*:565, 1984.

14

Rhythmic Excitation of the Heart

The heart is endowed with a special system (a) ır generating rhythmical impulses to cause ıythmical contraction of the heart muscle, and) for conducting these impulses rapidly throughıt the heart. When this system functions norıally, the atria contract about one sixth of a econd ahead of ventricular contraction, which llows extra filling of the ventricles before they ump the blood through the lungs and peripheral irculation. Another special importance of the sysem is that it allows all portions of the ventricles o contract almost simultaneously, which is essenial for effective pressure generation in the venricular chambers.

Unfortunately, though, this rhythmical and conluction system of the heart is very susceptible to lamage by heart disease, especially by ischemia f the heart tissues resulting from poor coronary lood flow. The consequence is often a very bizarre ıeart rhythm, or abnormal sequence of contraction f the heart chambers, and the pumping effectiveıess of the heart is often affected severely, even to he extent of causing death.

THE SPECIAL EXCITATORY AND CONDUCTIVE SYSTEM OF THE HEART

The adult human heart normally contracts at a rhythmic rate of about 72 beats per minute. Figure 14–1 illustrates the special excitatory and conductive system of the heart that controls these cardiac contractions. The figure shows (A) the *S-A node* in which the normal rhythmic self-excitatory impulse is generated, (B) the *internodal pathways* that conduct the impulse from the S-A node to the A-V node, (C) the *A-V node* in which the impulse from the atria is delayed before passing into the ventricles, (D) the *A-V bundle*, which conducts the impulse from the atria into the ventricles, and (E) the *left* and *right bundles of Purkinje fibers*, which conduct the cardiac impulse to all parts of the ventricles.

THE SINOATRIAL NODE

The sinoatrial (S-A) node is a small, flattened, ellipsoid strip of specialized muscle approximately 3 mm wide, 15 mm long, and 1 mm thick; it is located in the anterosuperior wall of the right atrium immediately anterior and lateral to the opening of the superior vena cava. The fibers of this node are each 3 to 5 microns in diameter, in contrast to a diameter of 15 to 20 microns for the surrounding atrial muscle fibers. However, the S-A fibers are continuous with the atrial fibers so that any action potential that begins in the S-A node spreads immediately into the atria.

Automatic Rhythmicity of the Sinoatrial Fibers

Most cardiac fibers have the capability of *self-excitation*, a process that can cause automatic rhythmical contraction. This is especially true of

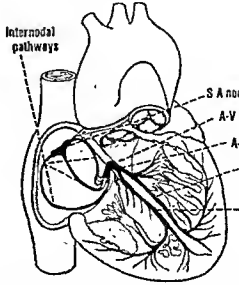


Figure 14–1. The S-A node and the Purkinje system of the heart, showing also the A-V node, the atrial internodal pathways, and the ventricular bundle branches.

the fibers of the heart's specialized conducting system; the portion of this system that displays self-excitation to the greatest extent is the fibers of the S-A node. For this reason, the sinoatrial node ordinarily controls the rate of beat of the entire heart, as will be discussed in detail later in this chapter. First, however, let us describe this automatic rhythmicity.

Mechanism of S-A Nodal Rhythmicity. The basic biophysics of self-excitatory rhythmicity of excitable tissues was presented in Chapter 10. However, let us review these principles as they apply to the S-A node.

Figure 14–2 illustrates action potentials recorded from an S-A nodal fiber for three heartbeats and, by comparison, a single ventricular muscle fiber action potential, shown to the right. Note that the "resting" potential of the S-A nodal fiber has a maximum negativity of only −55 to −60 millivolts in comparison with −85 to −90 millivolts for the ventricular fiber.

Before attempting to explain the rhythmicity of the S-A nodal fibers, first recall from the discussions of Chapters 10 and 13 that in cardiac muscle three different types of membrane ion channels play important roles in causing the voltage changes of the action potential. These are (1) the *fast sodium channels*, (2) the *slow calcium-sodium channels*, and (3) the *potassium channels*. The opening of the fast sodium channels for a few ten thousandths of a second is responsible for the very rapid spikelike onset of the action potential observed in ventricular muscle because of rapid influx of positive sodium ions to the interior of the fiber. Then the plateau of the ventricular action potential is caused, primarily by slower opening of the slow calcium-sodium channels, which lasts for a few tenths of a second. Finally, increased opening of the potassium channels and diffusion of large amounts of positive potassium ions out of the fiber return the membrane potential to its resting level.

But there is a difference in the function of these channels in the S-A nodal fiber. The cause of this difference is the much lesser negativity of the S-A nodal fiber during the "resting" potential, −55 millivolts. At this level of negativity, the fast sodium channels have mainly become "inactivated," which means that they have become blocked. The cause of this is that any time the membrane potential remains less negative than about −60 millivolts for a prolonged period of time, the gates on the inside of the cell membrane that close these channels become closed and remain so. Therefore, only the slow calcium-sodium channels can open (that is, can become "activated") and can thereby cause the action potential. Therefore, the action potential is slower to develop than that of the ventricular muscle and also recovers with a slow decrement of the potential rather than the abrupt recovery that occurs for the ventricular fiber.

However, there is still another feature of the S-A nodal fibers that is essential for their rhythmical self-excitation. This is the fact that they are quite leaky to sodium ions. This leakiness makes the "resting" potential of the S-A nodal fibers less negative than the resting potential of the ventricular muscle fibers, because entry of the positive sodium ions to the interior of the fiber neutralizes much of the fiber negativity. This leakiness of sodium ions also causes the "resting" potential gradually to rise between each two heart beats, as illustrated in Figure 14–2, until it finally reaches the *threshold voltage* of about −40 millivolts. At this point, this rising voltage suddenly opens the calcium-sodium channels, thus leading to the action potential. Therefore, basically, the inherent leakiness of the S-A nodal fibers to sodium ions causes their self-excitation.

Next, we must answer why this leakiness of sodium ions does not cause the S-A nodal fibers to remain depolarized all the time. The answer is this is that at the termination of the action potential, greatly increased numbers of potassium channels have become opened, and these allow far greater quantities of potassium to diffuse out of the fiber than normally. This carries a great excess of positive charges to the exterior and temporarily causes excess negativity inside the fiber, which is called a state of *hyperpolarization*. This hyperpolarization initially carries the "resting" membrane potential down to about −55 to −60 millivolts at the termination of the action potential.

Lastly, we must explain why the state of hyperpolarization also is not maintained forever. The reason is that during the next few tenths of a second after the action potential is over, progressively more and more of the potassium channels begin to close. Now the inward-leaking positive sodium ions once again overbalance the outward flux of potassium ions, which causes the "resting" potential to drift upward, finally reaching the threshold level for discharge at a potential of about −40 millivolts. Then the entire process begins

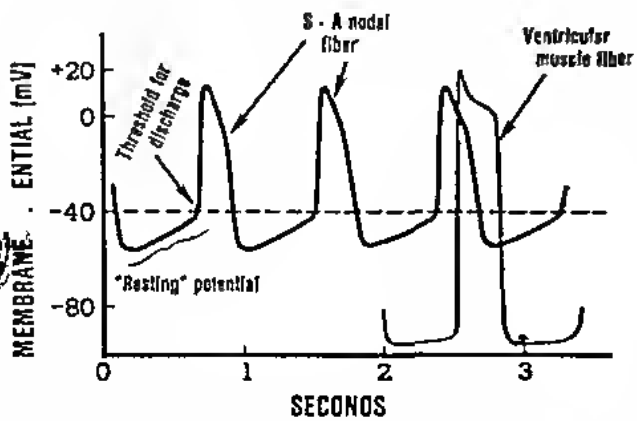


Figure 14–2. Rhythmic discharge of an S-A nodal fiber, and comparison of the S-A nodal action potential with that of a ventricular muscle fiber

again: self-excitation, recovery from the action potential, hyperpolarization after the action potential is over, upward drift of the "resting" potential, then re-excitation still again to elicit still another cycle; this process continues indefinitely throughout the life of the person.

INTERNODAL PATHWAYS AND TRANSMISSION OF THE CARDIAC IMPULSE THROUGH THE ATRIA

The ends of the S-A nodal fibers fuse with the surrounding atrial muscle fibers, and action potentials originating in the S-A node travel outward into these fibers. In this way, the action potential spreads through the entire atrial muscle mass and eventually also to the A-V node. The velocity of conduction in the atrial muscle is approximately 0.3 meter per second. However, conduction is somewhat more rapid in several small bundles of atrial muscle fibers. One of these, called the *anterior interatrial band*, passes through the anterior walls of the atria to the left atrium and conducts the cardiac impulse at a velocity of about 1 meter per second. In addition, three other small bundles curve through the atrial walls and terminate in the A-V node, also conducting the cardiac impulse at this rapid velocity. These three small bundles are illustrated in Figure 14–1 and are called respectively the *anterior, middle,* and *posterior internodal pathways*. The cause of the more rapid velocity of conduction in these bundles is the presence of a number of specialized conduction fibers mixed with the atrial muscle. These fibers are similar to the very rapidly conducting Purkinje fibers of the ventricles, which will be discussed subsequently.

THE ATRIOVENTRICULAR (A-V) NODE AND THE PURKINJE SYSTEM

Delay in Transmission at the A-V Node. Fortunately, the conductive system is organized so that the cardiac impulse will not travel from the atria into the ventricles too rapidly; this allows time for the atria to empty their contents into the ventricles before ventricular contraction begins. It is primarily the A-V node and its associated conductive fibers that delay this transmission of the cardiac impulse from the atria into the ventricles.

The A-V node is located in the septal wall of the right atrium immediately posterior to the tricuspid valve, as illustrated in Figure 14–1. Figure 14–3 shows diagrammatically the different parts of this node and its connections with the atrial internodal pathway fibers and the A-V bundle. The figure also shows the approximate intervals of time in fractions of a second between the genesis of the cardiac impulse in the S-A node and its appearance at different points in the A-V nodal system. Note that the impulse, after traveling through the internodal pathway, reaches the A-V node approximately 0.04 second after its origin in the S-A node. However, between this time and the time that the impulse emerges in the A-V bundle, another 0.11 second elapses. About one half of this time lapse occurs in the *junctional fibers*, which are very small fibers that connect the normal atrial fibers with the fibers of the node itself (illustrated in Figures 14–3 and 14–4). The velocity of conduction in these fibers is as little as 0.02 meter per second (about one twentieth that in normal cardiac muscle), which greatly delays entrance of the impulse into the A-V node. After entering the node proper, the velocity of conduction in the *nodal fibers* is still quite low, only 0.1 meter per second, about

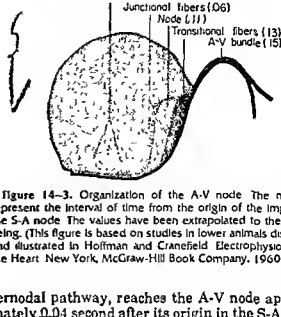


Figure 14–3. Organization of the A-V node. The numbers represent the interval of time from the origin of the impulse in the S-A node. The values have been extrapolated to the human being. (This figure is based on studies in lower animals discussed and illustrated in Hoffman and Cranefield Electrophysiology of the Heart. New York, McGraw-Hill Book Company, 1960.)

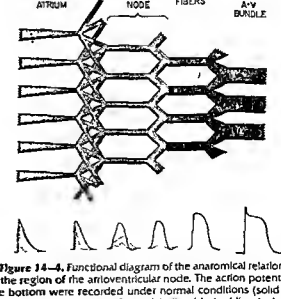


Figure 14–4. Functional diagram of the anatomical relationships in the region of the atrioventricular node. The action potentials at the bottom were recorded under normal conditions (solid lines) and under the influence of acetylcholine (dashed lines), showing blockage of conduction by this substance. (From Hoffman and Cranefield Electrophysiology of the Heart. New York, McGraw-Hill Book Company, 1960.)

one-fourth the conduction velocity in normal cardiac muscle. Therefore, a further delay in transmission occurs as the impulse travels through the A-V node into the *transitional fibers* and finally into the *A-V bundle* (also called the *bundle of His*).

Figure 14–4 illustrates a functional diagram of the A-V nodal region, showing the atrial fibers leading into the minute junctional fibers, then progressive enlargement of the fibers again as they spread through the node, through the transitional region, and into the A-V bundle.

Cause of the Delay. The cause of the extremely slow conduction in the junctional and some other A-V nodal fibers is partly their small size. The resistance to conduction of ionic currents both inside and outside small fibers is far greater in relation to the surface membrane area than is true for large fibers, and this makes it a general characteristic of small fibers that they conduct impulses at a slower velocity than do large fibers.

However, another feature of conduction at the A-V node probably accounts for most of the slow conduction. This is the fact that the junctional and A-V nodal fibers, like the S-A nodal fibers, have resting membrane potentials that are much less negative than the normal resting potential of other cardiac muscle. Therefore, the junctional and A-V nodal fibers, also like the S-A nodal fibers, have many if not most of their fast sodium channels blocked, as explained earlier in the chapter for the S-A fibers. As a consequence, the action potentials of these fibers must be initiated by the slow calcium-sodium channels, which are slow to open (to "activate") at the onset of the action potential, causing slow development of the action potential itself. Therefore, each linear segment of these fibers is also slow to excite the next segment, and this slowness continues throughout the length of the junctional and nodal fibers.

Prolonged Refractory Period of A-V Node. Another characteristic of the junctional and A-V nodal fibers is that their refractory period is about 25 per cent longer than the refractory period of usual cardiac muscle. The reason for this is probably the fact that the slow calcium-sodium channels that cause the action potential in these fibers close very slowly, thereby greatly prolonging the refractory period.

The importance of this prolonged refractory period is that it helps prevent the development of potentially serious arrhythmias of the heart that might occur if a second cardiac impulse were transmitted into the ventricular muscle too soon after the first, as we shall explain more fully later in the chapter.

TRANSMISSION IN THE PURKINJE SYSTEM

The *Purkinje fibers* that lead from the A-V node through the A-V bundle and into the ventricles have functional characteristics quite the opposite of those of the A-V nodal fibers; they are very large fibers, even larger than the normal ventricular muscle fibers, and they transmit impulses at a velocity of 1.5 to 4.0 meters per second, a velocity about 6 times that in the usual cardiac muscle and 150 times that in the junctional fibers. This allows almost immediate transmission of the cardiac impulse throughout the entire ventricular system.

The very rapid transmission of action potentials by Purkinje fibers is probably caused by increased numbers of nexuses between the successive cardiac cells that make up the Purkinje fibers. At these nexuses, ions are transmitted easily from one cell to the next, thus enhancing the velocity of transmission. The Purkinje fibers also have very few myofibrils, which means that they barely contract during the course of impulse transmission.

Distribution of the Purkinje Fibers in the Ventricles. The Purkinje fibers, after originating in the A-V node, form the A-V bundle, which then threads through the fibrous tissue between the valves of the heart and thence into the ventricular septum, as shown in Figure 14–1. The A-V bundle divides almost immediately into the *left* and *right bundle branches* that lie beneath the endocardium of the respective sides of the septum. Each of these branches spreads downward toward the apex of the respective ventricle, but also divides into small branches which spread around each ventricular chamber and finally back toward the base of the heart. The terminal Purkinje fibers penetrate about one third of the way into the muscle mass to terminate on the muscle fibers.

From the time that the cardiac impulse first enters the A-V bundle until it reaches the terminations of the Purkinje fibers, the total time that elapses averages about 0.03 second; therefore, once a cardiac impulse enters the Purkinje system, it spreads almost immediately to the entire endocardial surface of the ventricular muscle.

TRANSMISSION OF THE CARDIAC IMPULSE IN THE VENTRICULAR MUSCLE

Once the cardiac impulse has reached the ends of the Purkinje fibers, it is then transmitted through the ventricular muscle mass by the ventricular muscle fibers themselves. The velocity of transmission is now only 0.3 to 0.5 meter per second, one sixth that in the Purkinje fibers.

The cardiac muscle wraps around the heart in a double spiral with fibrous septa between the spiralling layers; therefore, the cardiac impulse does not necessarily travel directly outward toward the surface of the heart but instead angulates toward the surface along the directions of the spirals. Because of this, transmission from the endocardial surface to the epicardial surface of the ventricle requires as much as another 0.03 second, approximately equal to the time required for transmission through the entire Purkinje system. Thus, the total

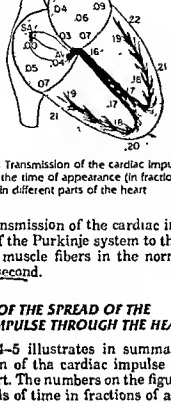


Figure 14–5. Transmission of the cardiac impulse through the heart, showing the time of appearance (in fractions of a second) of the impulse in different parts of the heart

time for transmission of the cardiac impulse from the origin of the Purkinje system to the last of the ventricular muscle fibers in the normal heart is about 0.06 second.

SUMMARY OF THE SPREAD OF THE CARDIAC IMPULSE THROUGH THE HEART

Figure 14–5 illustrates in summary form the transmission of the cardiac impulse through the human heart. The numbers on the figure represent the intervals of time in fractions of a second that lapse between the origin of the cardiac impulse in the S-A node and its appearance at each respective point in the heart. Note that the impulse spreads at moderate velocity through the atria but is delayed more than 0.1 second in the A-V nodal region before appearing in the A-V bundle. Once it has entered the bundle, it spreads rapidly through the Purkinje fibers to the entire endocardial surfaces of the ventricles. Then the impulse spreads slowly through the ventricular muscle to the epicardial surfaces.

It is extremely important that the student learn in detail the course of the cardiac impulse through the heart and the times of its appearance in each separate part of the heart, for a quantitative knowledge of this process is essential to the understanding of electrocardiography, which is discussed in the following three chapters.

CONTROL OF EXCITATION AND CONDUCTION IN THE HEART

THE S-A NODE AS THE PACEMAKER OF THE HEART

In the above discussion of the genesis and transmission of the cardiac impulse through the heart, it was stated that the impulse normally arises in the S-A node. However, this need not be the case under abnormal conditions, for other parts of the heart can exhibit rhythmic contraction in the same way that the fibers of the S-A node can; this is particularly true of the A-V nodal and Purkinje fibers.

The A-V nodal fibers, when not stimulated from some outside source, discharge at an intrinsic rhythmic rate of 40 to 60 times per minute, and the Purkinje fibers discharge at a rate of somewhere between 15 and 40 times per minute. These rates are in contrast to the normal rate of the S-A node of 70 to 80 times per minute.

Therefore, the question that we must ask is: Why does the S-A node control the heart's rhythmicity rather than the A-V node or the Purkinje fibers? The answer to this is simply that the rate of the S-A node is considerably greater than that of either the A-V node or the Purkinje fibers. Each time the S-A node discharges, its impulse is conducted into both the A-V node and the Purkinje fibers, discharging their excitable membranes. Then these tissues, as well as the S-A node, recover from the action potential and become hyperpolarized. But the S-A node loses this hyperpolarization much more rapidly than does either of the other two and emits a new impulse before either one of them can reach its own threshold for self-excitation. The new impulse again discharges both the A-V node and Purkinje fibers. This process continues on and on, the S-A node always exciting these other potentially self-excitatory tissues before self-excitation can actually occur.

Thus, the S-A node controls the beat of the heart, because its rate of rhythmic discharge is greater than that of any other part of the heart. Therefore, the S-A node is the normal *pacemaker* of the heart.

Abnormal Pacemakers—The Ectopic Pacemaker. Occasionally some other part of the heart develops a rhythmic discharge rate that is more rapid than that of the S-A node. For instance, this often occurs in the A-V node or in the Purkinje fibers. In either of these cases, the pacemaker of the heart shifts from the S-A node to the A-V node or to the excitable Purkinje fibers. Under rare conditions a point in the atrial or ventricular muscle develops excessive excitability and becomes the pacemaker.

A pacemaker elsewhere than the S-A node is called an *ectopic pacemaker*. Obviously, an ectopic pacemaker causes an abnormal sequence of contraction of the different parts of the heart.

Another cause of shift of the pacemaker is blockage of transmission of the impulses from the S-A node to the other parts of the heart, this occurring most frequently at the A-V node on the way to the ventricles. When A-V block occurs, the atria continue to beat at the normal rate of rhythm of the S-A node, while a new pacemaker develops in the Purkinje system of the ventricles and drives the ventricular muscle at a new rate somewhere between 15 and 40 beats per minute. However, the Purkinje system does not begin to emit its rhythmical impulses until 5 to 30 seconds later. During this entire period of time the ventricles fail to

pump blood, and the person faints after the first 4 to 5 seconds because of lack of blood flow to the brain This condition is called *Stokes-Adams syndrome*. If the period is too long, it can lead to death.

The cause of this prolonged delay before establishing the new rhythm is the phenomenon called *overdrive suppression* That is, when the S-A node drives the Purkinje system at a rhythmical frequency far above its natural frequency, as is normally the case, the Purkinje system is said to be overdriven, and for reasons not totally understood this temporarily suppresses the excitability of the Purkinje fibers so that an extra period of time is required before they can become self-excitatory. This is a very valuable effect because it ordinarily keeps the other parts of the heart besides the S-A node suppressed enough so that it is difficult for them to take over a pacemaker function.

ROLE OF THE PURKINJE SYSTEM IN CAUSING SYNCHRONOUS CONTRACTION OF THE VENTRICULAR MUSCLE

It is clear from the previous description of the Purkinja system that the cardiac impulse arrives at almost all portions of the ventricles within a very narrow span of time, exciting the first ventricular muscle fiber only 0.06 second ahead of excitation of the last ventricular muscle fiber Since the ventricular muscle fibers normally remain contracted for a total period of 0.3 second, one can see that this rapid spread of excitation throughout the entire ventricular muscle mass causes all portions of the ventricular muscle in both ventricles to contract at almost exactly the same time Effective pumping by the two ventricular chambers requires this synchronous type of contraction If the cardiac impulse traveled through the ventricular muscle very slowly, then much of the ventricular mass would contract prior to contraction of the remainder, in which case the overall pumping effect would be greatly depressed Indeed, in some types of cardiac debilities, some of which will be discussed in Chapters 16 and 17, such slow transmission does indeed occur, and the pumping effectiveness of the ventricles is decreased perhaps as much as 20 to 30 per cent.

CONTROL OF HEART RHYTHMICITY AND CONDUCTION BY THE AUTONOMIC NERVES

The heart is supplied with both sympathetic and parasympathetic nerves, as illustrated in Figure 13–11 of the previous chapter. The parasympathetic nerves are distributed mainly to the S-A and A-V nodes, to a lesser extent to the muscle of the two atria, and even less to the ventricular muscle. The sympathetic nerves, on the other are distributed to all parts of the heart, with a strong representation to the ventricular muscle as well as to all the other areas.

Effect of Parasympathetic (Vagal) Stimulation on Cardiac Rhythm and Conduction—Ventricular Escape. Stimulation of the parasympathetic nerves to the heart (the vagi) causes the hormone acetylcholine to be released at the vagal endings This hormone has two major effects on the heart. First, it decreases the rate of rhythm of the S-A node, and, second, it decreases the excitability of the A-V junctional fibers between the atrial musculature and the A-V node, thereby slowing transmission of the cardiac impulse into the ventricles. Very strong stimulation of the vagi can completely stop the rhythmic contraction of the S-A node or completely block transmission of the cardiac impulse through the A-V junction. In either case, rhythmic impulses are no longer transmitted into the ventricles. The ventricles stop beating for 4 to 10 seconds, but then some point in the Purkinje fibers, usually in the A-V bundle, develops a rhythm of its own and causes ventricular contraction at a rate of 15 to 40 beats per minute. This phenomenon is called *ventricular escape*.

Mechanism of the Vagal Effects. The acetylcholine released at the vagal nerve endings greatly increases the permeability of the fiber membranes to potassium, which allows rapid leakage of potassium to the exterior. This causes increased negativity inside the fibers, an effect called *hyperpolarization*, which makes excitable tissue much less excitable, as was explained in Chapter 10.

In the S-A node, the state of hyperpolarization decreases the "resting" membrane potential of the S-A nodal fibers to a level considerably more negative than the normal value, to a level as low as −65 to −75 millivolts rather than the normal level of −55 to −60 millivolts. Therefore, the upward drift of the resting membrane potential caused by sodium leakage requires much longer to reach the threshold potential for excitation. Obviously, this greatly slows the rate of rhythmicity of these nodal fibers. And, if the vagal stimulation is strong enough, it is possible to stop completely the rhythmical self-excitation of this node.

In the A-V node, the state of hyperpolarization makes it difficult for the minute junctional fibers, which can generate only small quantities of current during the action potential, to excite the nodal fibers. Therefore, the *safety factor* for transmission of the cardiac impulse through the junctional fibers and into the nodal fibers decreases. A moderate decrease in this simply delays conduction of the impulse, but a decrease in safety factor below unity (which means so low that the action potential of one fiber cannot cause an action potential in the successive fiber) completely blocks conduction.

Effect of Sympathetic Stimulation on Cardiac Rhythm and Conduction. Sympathetic stimulation causes essentially the opposite effects

on the heart to those caused by vagal stimulation as follows: First, it increases the rate of S-A nodal discharge. Second, it increases the rate of conduction and the excitability in all portions of the heart. Third, it increases greatly the force of contraction of all the cardiac musculature, both atrial and ventricular, as discussed in the previous chapter.

In short, sympathetic stimulation increases the overall activity of the heart. Maximal stimulation can almost triple the rate of heartbeat and can increase the strength of heart contraction as much as two-fold.

Mechanism of the Sympathetic Effect. Stimulation of the sympathetic nerves releases the hormone norepinephrine at the sympathetic nerve endings. The precise mechanism by which this hormone acts on cardiac muscle fibers is still somewhat doubtful, but the present belief is that it increases the permeability of the fiber membrane to sodium and calcium. In the S-A node, an increase of sodium permeability would cause increased tendency for the resting membrane potential to drift upward to the threshold level for self-excitation, which obviously would accelerate the onset of self-excitation after each successive heartbeat and therefore increase the heart rate.

In the A-V node, increased sodium permeability would make it easier for each fiber to excite the succeeding fiber, thereby decreasing the conduction time from the atria to the ventricles.

The increase in permeability to calcium ions is at least partially responsible for the increase in contractile strength of the cardiac muscle under the influence of sympathetic stimulation, because calcium ions play a powerful role in exciting the contractile process of the myofibrils.

ABNORMAL RHYTHMS OF THE HEART

Abnormal cardiac rhythms can be caused by (1) abnormal rhythmicity of the pacemaker itself, (2) shift of the pacemaker from the S-A node to other parts of the heart, (3) blocks at different points in the transmission of the impulse through the heart, (4) abnormal pathways of impulse transmission through the heart, and (5) spontaneous generation of abnormal impulses in almost any part of the heart. Some of these will be discussed in Chapter 17 in relation to electrocardiographic analysis of cardiac arrhythmias, but the major disturbances and their causes are presented here to illustrate some aberrations that can occur in the function of the rhythmicity and conducting systems of the heart.

HEART BLOCK

Occasionally, transmission of the impulse through the heart is blocked at a critical point in the conductive system. One of the most common of these points is between the atria and the ventricles; this condition is called *atrioventricular block*. Another common point is in one of the *bundle branches* of the Purkinje system. Rarely a block also develops between the S-A node and the atrial musculature.

Atrioventricular Block. In the human being, a block between the atria and the ventricles can result from localized damage or depression of the *A-V node*, especially of the *junctional fibers*, or of the *A-V bundle*. The causes include different types of infectious processes, excessive stimulation by the vagus nerves (which depresses conductivity in the junctional fibers), localized destruction of the A-V bundle as a result of a coronary infarct, pressure on the A-V bundle by arteriosclerotic plaques, or depression caused by various drugs.

Figure 14-6 illustrates a typical record of atrial and ventricular contraction while the A-V bundle was being progressively compressed to cause successive stages of block. During the first three contractions of the record, ventricular contraction followed in orderly sequence approximately 0.16 second after atrial contraction. Then, A-V bundle compression was begun, and the interval of time between the beginning of atrial contraction and the beginning of ventricular contraction increased steadily during the next five heartbeats from 0.16 to 0.32 second. Beyond this point further compression completely blocked impulse transmission. Thereafter, the atria continued to beat at their normal rate of rhythm, while the ventricles failed to contract at all for approximately 7 seconds. Then, "ventricular escape" occurred, and a rhythmic focus in the ventricular Purkinje system suddenly began to act as the ventricular pacemaker, causing ventricular contractions at a rate of approximately 30 per minute; these were completely dissociated from the atrial contractions. These various stages of heart block will be discussed more fully in Chapter 17, along with presentation of the electrocardiographic abnormalities caused by the block.

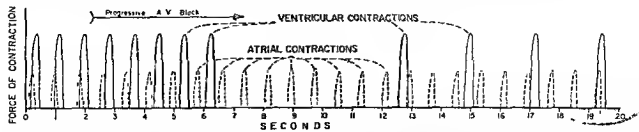


Figure 14-6. Contraction of the atria and ventricles of a heart, showing the effect of progressive A-V block. Note the progressive increase in the interval between the onset of atrial contraction and the onset of ventricular contraction. Note also the final complete asynchronism between atrial and ventricular contraction

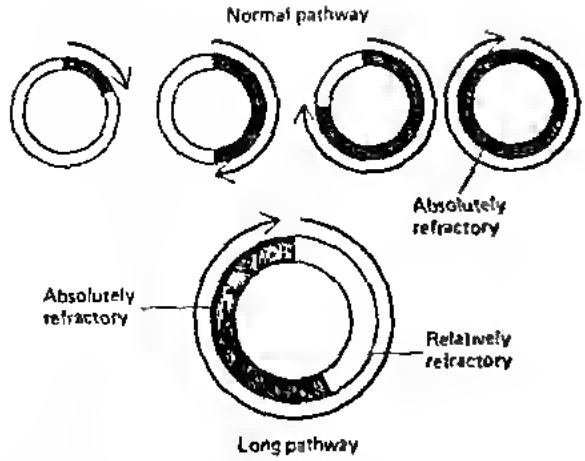


Figure 14–7. The circus movement showing annihilation of the impulse in the short pathway and continued propagation of the impulse in the long pathway

THE PHENOMENON OF "RE-ENTRY" AND ABNORMAL RHYTHMS

Normally, when the cardiac impulse has traveled throughout the entire extent of the heart, it has no place else to go and therefore it simply dies. Then the heart remains quiescent until a new signal begins in the S-A node.

However, under certain conditions this normal sequence of events does not occur. Instead, the cardiac impulse sometimes travels around and around in cardiac muscle without stopping. This is the phenomenon called re-entry, also called the circus movement. It is a totally abnormal situation, but it can cause very serious cardiac arrhythmias, some of which are lethal. Therefore, let us explain more fully the cause of re-entry.

Figure 14–7 illustrates several small cardiac muscle strips cut in the form of circles. If such a strip is stimulated at the 12 o'clock position so *that the impulse travels in only one direction*, the impulse spreads progressively around the circle until it returns to the 12 o'clock position. If the originally stimulated muscle fibers are still in a refractory state, the impulse then dies out, for refractory muscle cannot transmit a second impulse. However, there are three different conditions that can cause this impulse to continue to travel around the circle, that is, to cause "re-entry" of the impulse into muscle that has already been excited.

First, if the *length of the pathway around the circle is long*, by the time the impulse returns to the 12 o'clock position the originally stimulated muscle will no longer be refractory, and the impulse will continue around the circle again and again.

Second, if the length of the pathway remains constant but the *velocity of conduction becomes decreased* enough, an increased interval of time will elapse before the impulse returns to the 12 o'clock position. By this time the originally stimulated muscle might be out of the refractory state, and the impulse can continue around the circle again and again.

Third, *the refractory period of the muscle might become greatly shortened*. In this case, the impulse could also continue around and around the circle.

All three of these conditions occur in different pathoal states of the human heart as follows: (1) A long pathway frequently occurs in dilated hearts. (2) Decreased rate of conduction frequently results from blockage of the Purkinje system, ischemia of the muscle, high blood potassium, and many other factors. (3) A shortened refractory period frequently occurs in response to various drugs, such as epinephrine, or following repetitive electrical stimulation. Thus, in many different cardiac disturbances re-entry can cause abnormal patterns of cardiac contraction or abnormal cardiac rhythms that completely ignore the pace-setting effects of the S-A node.

Two especially serious rhythms caused by re-entrant impulses are *flutter* and *fibrillation*. Flutter means extremely rapid heart beat at rates sometimes as high as 200 to 350 beats per minute but with reasonably coordinated contractions of the cardiac muscle. This type of rhythm occurs frequently in the atria but only rarely in the ventricles. Fibrillation is different from flutter in that it has a very high frequency and the contractions are incoordinate, so much so that almost no portions of the cardiac muscle contract in unison, and therefore the atria or ventricles fail to pump significant amounts of blood. For this reason, ventricular fibrillation is lethal, as we shall discuss more fully subsequently.

Atrial Flutter Resulting from a Circus Pathway

The left-hand panel of Figure 14–8 illustrates a circus pathway around and around the atria from top to bottom, passing to the left of the superior and inferior venae cavae. Such circus pathways have been initiated experimentally in the atria of dogs' hearts, and electrocardiographic records, which will be discussed in Chapter 17, indicate that this type of circus pathway also develops in the human heart when the atria become greatly dilated as a result of valvular heart disease. The rate of flutter is usually 200 to 350 times per minute.

Partial Block at the A-V Node During Atrial Flutter. The refractory period of the Purkinje fibers and ventricular muscle is approximately 1/200 minute so that not over 200 impulses per minute usually can be transmitted into the ventricles. Therefore, when the atrium contracts as rapidly as 300 times per minute, only one of every two impulses passes into the ventricles, thus causing the atria to beat at a rate two times that of the ventricles. Occasionally, a 3:1 or, rarely, a 4:1 rhythm of the heart develops because overactivity in the A-V node suppresses its conductivity and reduces

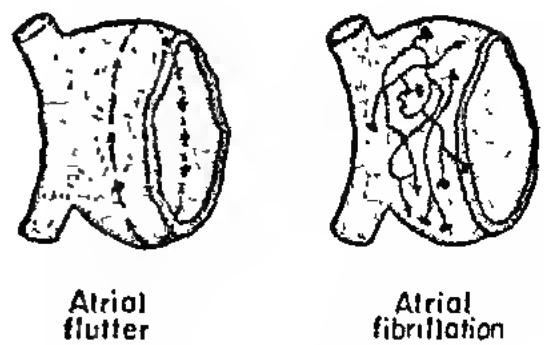


Figure 14–8. Pathways of impulses in atrial flutter and atrial fibrillation.

the number of impulses conducted into the ventricles even lower.

The "Chain Reaction" Mechanism of Fibrillation

Fibrillation, whether it occurs in the atria or in the ventricles, is a very different condition from flutter. One can see many separate contractile wavelets spreading in different directions over the cardiac muscle at the same time in either atrial or ventricular fibrillation. Obviously, then, the re-entrant impulses in fibrillation are entirely different from those in flutter. One of the best ways to explain the mechanism of fibrillation is to describe the initiation of fibrillation by stimulation with 60 cycle alternating electrical current.

Fibrillation Caused by 60 Cycle Alternating Current. At a central point in the ventricles of heart A in Figure 14–9, a 60 cycle electrical stimulus is applied through a stimulating electrode. The first cycle of the electrical stimulus causes a depolarization wave to spread in all directions, leaving all the muscle beneath the electrode in a refractory state After about 0 25 second, this muscle begins to come out of the refractory state. However, some portions of the muscle come out of refractoriness prior to other portions This state of events is depicted in heart A by many light patches, which represent excitable cardiac muscle, and dark patches, which represent still refractory muscle. New stimuli from the electrode can now cause impulses to travel in certain directions through the heart but not in all directions. Thus, in heart A certain impulses travel for short distances until they reach refractory areas of the heart and then are blocked. Other impulses, however, pass between the refractory areas and continue to travel in the excitable patches of muscle. Now, several events transpire in rapid succession, all occurring simultaneously and eventuating in a state of fibrillation These are:

First, block of the impulses in some directions but successful transmission in other directions creates one of the necessary conditions for a re-entrant signal to develop—that is, *transmission of some of the depolarization waves around the heart in only one direction.* As a result, these waves do not run into waves traveling in the opposite direction and therefore do not annihilate themselves on the opposite side of the heart but can continue around and around the ventricles.

Second, the rapid stimulation of the heart causes two changes in the cardiac muscle itself, both of which predispose to circus movement: (1) The *velocity of conduction through the heart becomes decreased,* which allows a longer time interval for the impulses to travel around the heart. (2) The *refractory period of the muscle becomes shortened,* allowing re-entry of the impulse into previously excited heart muscle within a much shorter period of time than normally.

Third, one of the most important features of fibrillation is the *division of impulses,* as illustrated in heart A. When a depolarization wave reaches a refractory area in the heart, it travels to both sides around the area. Thus, a single impulse becomes two impulses. Then when each of these reaches another refractory area it, too, divides to form still two more impulses. In this way many different new wave fronts are continually being formed in the heart by a progressive *chain reaction* until, finally, there are many small depolarization waves traveling in many different directions at the same time Furthermore, this irregular pattern of impulse travel causes a *circuitous route for the impulses to travel, greatly lengthening the conductive pathway, which is one of the conditions leading to fibrillation.* It also results in a continual irregular pattern of patchy refractory areas in the heart. One can readily see that a vicious cycle has been initiated: more and more impulses are formed; these cause more and more patches of refractory muscle; and the refractory patches cause more and more division of the impulses Therefore, any time a single area of cardiac muscle comes out of refractoriness, an impulse is always close at hand to re-enter the area.

Heart B in Figure 14–9 illustrates the final state that develops in fibrillation. Here one can see many impulses traveling in all directions, some dividing and increasing the number of impulses while others are blocked entirely by refractory areas In the final state of fibrillation, the number of new impulses being formed exactly equals the number of impulses that are being blocked by refractory areas. Thus, a steady state has developed with a certain average number of impulses traveling all the time in all directions through the ventricular syncytium

As is the case with many other cardiac arrhythmias, fibrillation is usually confined to either the atria or the ventricles alone and not to both syncytial masses of muscle at the same time, because these two masses of muscles are electrically insulated from each other by the rings of fibrous tissue around the heart valves.

Vulnerable Period for Causing Cardiac Fibrillation. In the previous discussion of fibrillation caused by 60 cycle current, it was noted that when patchy areas of ventricular muscle are still refractory but are surrounded by other areas of muscle that have come out of refractoriness, each excitatory wave can then pass into the recovered ventricular muscle on the two sides of each refractory area, thus multiplying the number of excitatory waves. This obviously predisposes to fibrillation. The time during the heart cycle when there are likely to be simultaneous areas of refractoriness and nonrefractoriness in the heart muscle is at the very moment that the heart is recovering from the previous cardiac cycle—that is, just at the end of the cardiac action potential Therefore, this instant in the cardiac

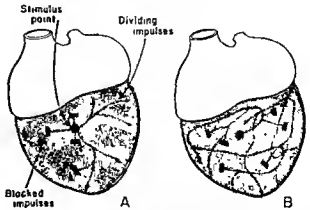


Figure 14–9. *A.* Initiation of fibrillation in a heart when patches of refractory musculature are present. *B.* Continued propagation of *fibrillatory impulses* in the fibrillating ventricle.

cycle is said to be the *vulnerable period* of the ventricles for fibrillation. In fact, a single electric shock during this vulnerable period frequently can lead to fibrillation.

Importance of the Purkinje System in Preventing Ventricular Fibrillation and Other Re-Entrant Phenomena. By now, it should be clear that the Purkinje system plays an exceedingly important role in preventing ventricular fibrillation and other re-entrant rhythms. There are two major reasons for this.

First, the normal Purkinje system conducts the cardiac impulse to all parts of the heart within a few hundredths of a second, long before any of the ventricular muscle can come out of refractoriness. Therefore, the action potentials in the ventricular muscle fail to propagate any further, and the cardiac impulse simply dies until a new S-A nodal impulse is generated.

Second, the Purkinje fibers, like the A-V nodal fibers, have a refractory period about 25 per cent greater than that of the ventricular muscle. Therefore, by the time the Purkinje fibers are no longer refractory the ventricular muscle fibers have already long been out of the refractory state. As a result, all portions of the ventricular muscle are already available to accept new action potentials from the Purkinje system when it is again excited. Therefore, this long refractory period of the Purkinje fibers in relation to the somewhat shorter refractory period of the ventricular muscle protects the muscle from being excited during its vulnerable period for fibrillation.

Atrial Fibrillation

Atrial fibrillation is completely different from atrial flutter because the circus movement does not travel in a regular pathway. Instead, many different excitation waves can be seen to travel over the surfaces of the atria at the same time. Atrial fibrillation occurs frequently when the atria become greatly overdilated—in fact, many times as frequently as flutter. When flutter does occur, it usually becomes fibrillation after a few days or weeks. To the right in Figure 14-8 are illustrated the pathways of fibrillatory impulses traveling through the atria.

Obviously, atrial fibrillation results in complete incoordination of atrial contraction so that atrial pumping ceases entirely.

Effect of Atrial Fibrillation on the Overall Pumping Effectiveness of the Heart. The normal function of the atria is to help fill the ventricles. However, the atria are probably responsible for not more than 20 to 30 per cent of the normal ventricular filling, which was explained in Chapter 13. Therefore, even when the atria fail to act as primer pumps because of atrial fibrillation, the ventricles can still fill enough so that the effectiveness of the heart as a pump is reduced only 20 to 30 per cent, which is well within the "cardiac reserve" of all but severely weakened hearts. For this reason, atrial fibrillation can continue for many years without serious cardiac debility.

Irregularity of the Ventricular Rhythm During Atrial Fibrillation. When the atria are fibrillating, impulses arrive at the A-V node rapidly but also irregularly. Since the A-V node will not pass a second impulse for approximately 0.35 second after a previous one, at least 0.35 second must elapse between one ventricular [illegible] d the next, and an additional but variable [illegible]rval of 0 to 0.6 second occurs before one of the [illegible]ar fibrillatory impulses happens to arrive at the A-V node. Thus, the interval between successive ventricular contractions varies from a minimum of about 0.35 second to a maximum of about 0.95 second, causing a very irregular heart beat. In fact, this irregularity is one of the clinical findings used to diagnose the condition.

Ventricular Fibrillation

Ventricular fibrillation is extremely important because at least one quarter of all persons die in ventricular fibrillation. For instance, the hearts of most patients with coronary infarcts fibrillate shortly before death. In only a few instances on record have fibrillating human ventricles been known to return of their own accord to a rhythmical beat.

The likelihood of the occurrence of re-entrant circus movements in the ventricles and, consequently, of ventricular fibrillation is greatly increased when the ventricles are dilated or when the rapidly conducting *Purkinje system is blocked* so that impulses cannot be transmitted rapidly. Also, *electric shock*, particularly with 60 cycle electric current, as discussed previously, or *rhythmical foci in ischemic muscle*, which will be discussed, are common initiating causes of ventricular fibrillation.

Inability of the Heart to Pump Blood During Ventricular Fibrillation. When the ventricles begin to fibrillate, the different parts of the ventricles no longer contract simultaneously. For the first few seconds the ventricular muscle undergoes rather coarse contractions which may pump a few milliliters of blood with each contraction. However, the impulses in the ventricles rapidly become divided into many much smaller impulses, and the contractions become so fine and asynchronous, rather than coarse, that they pump no blood whatsoever. The ventricles dilate because of failure to pump the blood that is flowing into them, and within 60 to 90 seconds the ventricular muscle becomes too weak, because of lack of coronary blood supply, to contract strongly even if coordinate contraction should return. Therefore, death is immediate when ventricular fibrillation begins.

Myocardial Ischemia as the Usual Cause of Ventricular Fibrillation. The most frequent cause of ventricular fibrillation is the development of a re-entrant circus movement adjacent to an area of myocardial ischemia caused by occlusion of a branch of a coronary artery. The cardiac muscle in the middle of the ischemic area frequently receives so little nutrition that it dies or becomes totally nonfunctional. However, the muscle along the edges of the ischemia can still conduct action potentials at a very slow velocity. This sets up a condition that predisposes to fibrillation. Re-entrant pathways frequently develop either wholly in the ischemic muscle or between this and the surrounding muscle, leading to a continuous circus movement that emits impulses in all directions away from the focus. The rapid repetitiveness of these impulses creates patchy areas of refractoriness in the adjacent ventricular muscle, and this is the essential condition for initiating the chain reaction mechanism of fibrillation, as discussed earlier. Thus, the fibrillatory state ensues in almost exactly the same manner as that described earlier for stimulation with a 60 cycle electrical curren[illegible] clear, therefore, that any focus that emits r[illegible] es is [illegible]ely to initiate fibrillation.

Electrical Defibrillation o[illegible] [illegible]hough

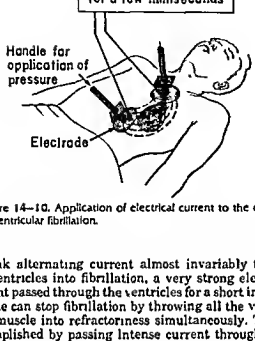


Figure 14–10. Application of electrical current to the chest to stop ventricular fibrillation.

a weak alternating current almost invariably throws the ventricles into fibrillation, a very strong electrical current passed through the ventricles for a short interval of time can stop fibrillation by throwing all the ventricular muscle into refractoriness simultaneously. This is accomplished by passing intense current through electrodes placed on two sides of the heart. The current penetrates most of the fibers of the ventricles, thus stimulating essentially all parts of the ventricles simultaneously and causing them to become refractory. All impulses stop, and the heart then remains quiescent for 3 to 5 seconds, after which it begins to beat again, with the S-A node or some other part of the heart becoming the pacemaker. Occasionally, however, the same re-entrant focus that had originally thrown the ventricles into fibrillation is still present, and fibrillation begins again immediately.

When electrodes are applied directly to the two sides of the heart, fibrillation can usually be stopped with 110 volts of 60 cycle alternating current applied for 0.1 second, or 1000 volts direct current applied for a few thousandths of a second. When applied through the chest wall, as illustrated in Figure 14–10, the usual procedure is to charge a large electrical capacitor up to several thousand volts and then cause the capacitor to discharge in a few thousandths of a second through the electrodes and the heart. In our laboratory the heart of a single anesthetized dog was defibrillated 130 times through the chest wall, and the animal remained in perfectly normal condition.

Hand Pumping of the Heart ("Cardiac Massage") As an Aid to Defibrillation

Unless defibrillated within one minute after fibrillation begins, the heart is usually too weak to be revived by defibrillation alone. However, it is still possible to revive the heart by preliminarily pumping it by hand and then defibrillating it later. In this way small quantities of blood are delivered into the aorta, and a renewed coronary blood supply develops. Then, after a few minutes, electrical defibrillation often becomes possible. Indeed, fibrillating hearts have been pumped by hand as long as 90 minutes before defibrillation. In recent years, a technique of pumping the heart without opening the chest has been developed; this technique consists of intermittent, very powerful thrusts of pressure on the chest wall.

Lack of blood flow to the brain for more than five to ten minutes usually results in permanent mental impairment or even total destruction of the brain. Even though the heart should be revived, the person might die from the effects of brain damage or live with permanent mental impairment.

PREMATURE CONTRACTION—ECTOPIC FOCI

Often, a small area of the heart becomes much more excitable than normal and causes an abnormal impulse to be generated during the time interval between the normal impulses. This can occur in either the atria or the ventricles. A depolarization wave spreads outward from the irritable area and initiates a *premature contraction* of the heart. The point at which the abnormal impulse is generated is called an *ectopic focus*.

A frequent cause of an ectopic focus is an irritable area in cardiac muscle that becomes self-excitable. This can result from a local area of muscle ischemia, overuse of stimulants such as caffeine or nicotine, lack of sleep, anxiety, or other debilitating states. However, in many instances, especially when there is ischemia, the ectopic focus is caused by a re-entrant signal that is delayed for a short period of time by slow transmission of the action potential in the ischemic muscle; then this signal re-enters the ventricular muscle a few tenths of a second later after the normal heart contraction has already occurred. This obviously will cause a second contraction of the heart at an abnormal time.

CARDIAC TACHYCARDIAS

Sometimes an ectopic focus becomes so irritable that it establishes a rhythmical contraction of its own at a more rapid rate than that of the S-A node. When this occurs, the ectopic focus becomes the pacemaker of the heart. In most instances of this, the repetitive rhythm is caused by re-entrant signals that have established small circus movements in the ectopic area.

A common point for development of an ectopic pacemaker is the A-V node itself, which can establish a re-entrant pathway within its own boundaries and cause a very rapid rate of heart contraction, as will be discussed in more detail in Chapter 17. Another frequent cause of rapid heart rate is an ectopic pacemaker in an ischemic area of ventricular muscle. Here again, the mechanism is usually a focal circus movement that transmits repetitive signals throughout the ventricle and causes *ventricular tachycardia*, which is also discussed more fully in Chapter 17.

CARDIAC ARREST

When cardiac metabolism becomes greatly disturbed as a result of any one of many possible conditions, the rhythmic contractions of the heart occasionally stop. One of the most common causes of cardiac arrest is hypoxia of the heart, often occurring when a patient is anesthetized too deeply or when the coronary blood flow to the S-A node is blocked. Severe hypoxia prevents the muscle fibers from maintaining normal ionic differen-

tials across their membranes, and the excitability may be so affected that the automatic rhythmicity disappears.

Occasionally, patients with severe myocardial disease develop cardiac arrest, which obviously can lead to death. In many cases, however, rhythmic electrical impulses from an implanted electronic cardiac "pacemaker" have been used successfully to keep patients alive for many years.

REFERENCES

Akera, T., and Brody, T. M. Myocardial membranes: Regulation and function of the sodium pump. *Annu. Rev. Physiol.* *44*:375, 1982.

Brown, H. F. Electrophysiology of the sinoatrial node. *Physiol. Rev.*, *62*:505, 1982.

Bruisaert, D. L., and Paulus, W. J. Contraction and relaxation of the heart as muscle and pump. *In* Guyton, A. C., and Young, D. B. (eds.) International Review of Physiology: Cardiovascular Physiology III. Vol. 18. Baltimore, University Park Press, 1979, p. 1.

Chung, E. K. (ed.) Artificial Cardiac Pacing: Practical Approach. Baltimore: Williams & Wilkins, 1978.

Cobb, L. A., et al. Community cardiopulmonary resuscitation. *Annu. Rev. Med.*, *31*:433, 1980.

Del Negro, A. A., and Fletcher, R. D. Indications for and Use of Artificial Cardiac Pacemakers. Chicago, Year Book Medical Publishers, 1978.

DiFrancesco, D., and Noble, D. A model of cardiac electrical activity incorporating ionic pumps and concentration changes. *Phil. Trans. R. Soc. Lond. (Biol.)* *307*:353, 1983.

Durrer, D., *et al.* Human cardiac electrophysiology. *In* Dickinson, C. J., and Marks, J. (eds.) Developments in Cardiovascular Medicine. Lancaster, England: MTP Press, 1978, p. 53.

Fliss, D. Na-Ca exchange in cardiac tissues. *Adv. Myocardiol.* *5*:295, 1985.

Farah, A. E., Alousi, A. A., and Schwartz, R. P., Jr. Positive inotropic agents. *Annu. Rev. Pharmacol. Toxicol.* *24*:275, 1984.

Fozzard, H. A. Heart: Excitation-contraction coupling. *Annu. Rev. Physiol.*, *39*:201, 1977.

Geddes, L. A. Cardiovascular Medical Devices. New York, John Wiley & Sons, 1984.

Gilmour, R. F., Jr., and Zipes, D. P. Slow inward current and cardiac arrhythmias. *Am. J. Cardiol.* *55*:89B, 1985.

Glitsch, H. G. Electrogenic Na pumping in the heart. *Annu. Rev. Physiol.*, *44*:389, 1982.

Guyton, A. C., and Satterfield, J. Factors concerned in electrical defibrillation of the heart, particularly through the unopened chest. *Am. J. Physiol.* *167*:81, 1951.

Herbette, L., *et al.* The interaction of drugs with the sarcoplasmic reticulum. *Annu. Rev. Pharmacol. Toxicol.*, *22*:413, 1982.

Hondeghem, L. M., and Katzung, B. G. Antiarrhythmic agents: The modulated receptor mechanism of action of sodium and calcium channel blocking drugs. *Annu. Rev. Pharmacol. Toxicol.*, *24*:387, 1984.

Irisawa, H. Comparative physiology of the cardiac pacemaker mechanism. *Physiol. Rev.* *58*:461, 1981.

Jacobson, L. Cardiac Pacing: Principles and Case Studies. New Hyde Park, New York, Medical Examination Publishing Co., 1981.

Jones, P. Cardiac Pacing. New York, Appleton Century Crofts, 1980.

Josephson, M. E., and Seides, S. F. Clinical Cardiac Electrophysiology, Techniques and Interpretations. Philadelphia, Lea & Febiger, 1979.

Josephson, M. E., and Singh, B. N. Use of calcium antagonists in ventricular dysfunction. *Am. J. Cardiol.*, *55*:81B, 1985.

Langer, G. A. Sodium-calcium exchange in the heart. *Annu. Rev. Physiol.*, *44*:435, 1982.

Latorre, R., *et al.* K^+ channels gated by voltage and ions. *Annu. Rev. Physiol.*, *46*:485, 1984.

Lazdunski, M., and Renaud, J. F. The action of cardiotoxins on cardiac plasma membranes. *Annu. Rev. Physiol.*, *44*:463, 1982.

Levy, M. N., *et al.* Neural regulation of the heart beat. *Annu. Rev. Physiol.*, *43*:443, 1981.

Levy, M. N., and Martin, P. J. Neural control of the heart. *In* Berne, R. M., *et al.* (eds.) Handbook of Physiology. Sec. 2, Vol. I. Baltimore, Williams & Wilkins, 1979, p. 581.

Loewenstein, W. R. Junctional intercellular communication: The cell-to-cell membrane channel. *Physiol. Rev.* *61*:829, 1981.

McAnulty, J., and Rahimtoola, S. Prognosis in bundle branch block. *Annu. Rev. Med.*, *32*:429, 1981.

McDonald, T. F. The slow inward calcium current in the heart. *Annu. Rev. Physiol.*, *44*:425, 1982.

Meijler, F. L. Atrioventricular conduction versus heart size from mouse to whale. *J. Am. Coll. Cardiol.*, *5*:363, 1985.

Nobel, D. The Initiation of the Heartbeat. New York, Oxford University Press, 1979.

Orrego, F. Calcium and the mechanism of action of digitalis. *Gen. Pharmacol.* *15*:273, 1984.

Pick, A., and Langendorf, R. Interpretation of Complex Arrhythmias. Philadelphia, Lea & Febiger, 1980.

Reuter, H. Properties of two inward membrane currents in the heart. *Annu. Rev. Physiol.* *41*:413, 1979.

Reuter, H. Ion channels in cardiac cell membranes. *Annu. Rev. Physiol.*, *44*:473, 1984.

Sanguinetti, M. C., and Kass, R. S. Regulation of cardiac calcium channel current and contractile activity by the dihydropyridine Bay K 8644 is voltage-dependent. *J. Mol. Cell. Cardiol.*, *16*:667, 1984.

Spear, J. F., and Moore, E. N. Mechanisms of cardiac arrhythmias. *Annu. Rev. Physiol.* *44*:485, 1982.

Sperelakis, N. Origin of the cardiac resting potential. *In* Berne, R. M., *et al.* (eds.) Handbook of Physiology. Sec. 2, Vol. I. Baltimore: Williams & Wilkins, 1979, p. 187.

Sperelakis, N. Propagation mechanisms in heart. *Annu. Rev. Physiol.* *41*:441, 1979.

Sperelakis, N. Hormonal and neurotransmitter regulation of Ca^+ influx through voltage-dependent slow channels in cardiac muscle membrane. *Membr. Biochem.*, *5*:131, 1984.

Stull, J. T., and Mayer, S. E. Biochemical mechanisms of adrenergic and cholinergic regulation of myocardial contractility. *In* Berne, R. M., *et al.* (eds.) Handbook of Physiology. Sec. 2, Vol. I. Baltimore, Williams & Wilkins, 1979, p. 741.

Vassalle, M. Electrogenesis of the plateau and pacemaker potential. *Annu. Rev. Physiol.*, *41*:425, 1979.

Verrier, R. L., and Lown, B. Behavioral stress and cardiac arrhythmias. *Annu. Rev. Physiol.* *46*:155, 1984.

15

The Normal Electrocardiogram

Transmission of the depolarization wave, also commonly called the *cardiac impulse*, through the heart has been discussed in detail in Chapter 14 As the wave passes through the heart, electrical currents spread into the tissues surrounding the heart, and a small proportion of these spreads all the way to the surface of the body. If electrodes are placed on the skin on opposite sides of the heart, electrical potentials generated by the heart can be recorded; the recording is known as an *electrocardiogram*. A normal electrocardiogram for two beats of the heart is illustrated in Figure 15–1

CHARACTERISTICS OF THE NORMAL ELECTROCARDIOGRAM

The normal electrocardiogram is composed of a P wave, a "QRS complex," and a T wave. The QRS complex is often three separate waves, the Q wave, the R wave, and the S wave.

The P wave is caused by electrical currents generated as the atria depolarize prior to contraction, and the QRS complex is caused by currents generated when the ventricles depolarize prior to contraction, that is, as the depolarization wave spreads through the ventricles. Therefore, both the P wave and the components of the QRS complex are *depolarization waves*. The T wave is caused by currents generated as the ventricles recover from the state of depolarization. This process occurs in ventricular muscle 0.25 to 0.35 second after depolarization, and this wave is known as a *repolarization wave*. Thus, the electrocardiogram is composed of both depolarization and repolarization waves. The principles of depolarization and repolarization were discussed in Chapter 10 However, the distinction between depolarization waves and repolarization waves is so important in electrocardiography that further clarification is needed, as follows:

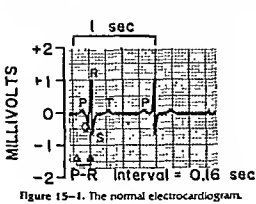


Figure 15–1. The normal electrocardiogram.

DEPOLARIZATION WAVES VERSUS REPOLARIZATION WAVES

Figure 15–2 illustrates a muscle fiber in four different stages of depolarization and repolarization. During the process of "depolarization" the normal negative potential inside the fiber is lost and the membrane potential actually reverses; that is, it becomes slightly positive inside and negative outside.

In Figure 15–2A the process of depolarization, illustrated by the red-colored positive charges inside and negative charges outside, is traveling from left to right, and the first half of the fiber has already depolarized while the remaining half is still polarized Therefore, the left electrode on the fiber is in an area of negativity where it touches the outside of the fiber, while the right electrode is in an area of positivity; this causes the meter to record positively. To the right of the muscle fiber is illustrated a record of the potential between the electrodes as recorded by a high-speed recording meter at this particular stage of depolarization. Note that when depolarization has reached this halfway mark, the record has risen to a maximum positive value.

In Figure 15–2B depolarization has extended over the entire muscle fiber, and the recording to the right has returned to the zero base line because both electrodes are now in areas of equal negativity. The completed wave is a *depolarization wave* because it results from spread of depolarization along the entire extent of the muscle fiber.

Figure 15–2C illustrates the repolarization process in the muscle fiber, which has proceeded halfway along the fiber from left to right. At this point, the left electrode is in an area of positivity, while the right electrode is in an area of negativity. This is opposite to the polarity in Figure 15–2A. Consequently, the recording, as illustrated to the right, becomes negative.

Finally, in Figure 15–2D the muscle fiber has completely repolarized, and both electrodes are now in areas of positivity so that no potential is recorded between them. Thus, in the recording to the right, the potential

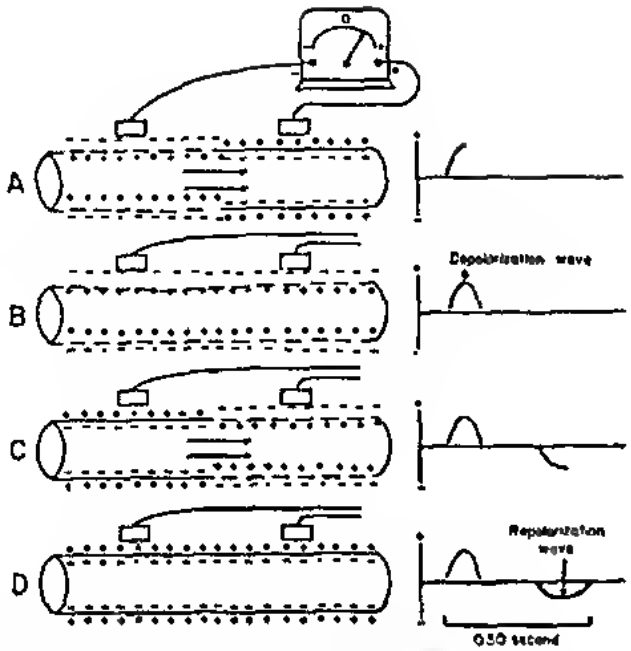


Figure 15–2. Recording the *depolarization wave* and the *repolarization wave* from a cardiac muscle fiber

returns once more to the zero level This completed negative wave is a *repolarization wave* because it results from spread of the repolarization process over the muscle fiber

Relationship of the Monophasic Action Potential of Cardiac Muscle to the QRS and T Waves. The monophasic action potential of ventricular muscle, which was discussed in the preceding chapter, normally lasts between 0.25 and 0 35 second The top part of Figure 15–3 illustrates a monophasic action potential recorded from a microelectrode inserted to the inside of a single ventricular muscle fiber The upsweep of this action potential is caused by *depolarization*, and the return of the potential to the base line is caused by *repolarization*

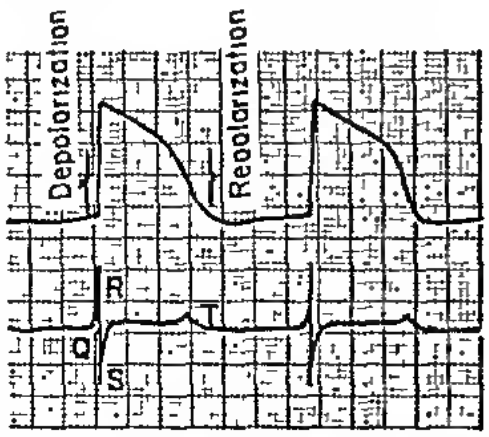


Figure 15–3. *Above* Monophasic action potential from a ventricular muscle fiber during normal cardiac function, showing rapid depolarization and then repolarization occurring slowly during the plateau stage but very rapidly toward the end. *Below* Electrocardiogram recorded simultaneously

Note below the simultaneous recording of the electrocardiogram from this same ventricle, which shows the QRS wave appearing at the beginning of the monophasic action potential and the T wave appearing at the end. Note especially that *no potential at all is recorded in the electrocardiogram when the ventricular muscle is either completely polarized or completely depolarized*. It is only when the muscle is partly polarized and partly depolarized that current flows from one part of the ventricles to another part and therefore also flows to the surface of the body to cause the electrocardiogram

RELATIONSHIP OF ATRIAL AND VENTRICULAR CONTRACTION TO THE WAVES OF THE ELECTROCARDIOGRAM

Before contraction of muscle can occur, depolarization must spread through the muscle to initiate the chemical processes of contraction The P wave results from spread of depolarization through the atria, and the QRS wave from spread of depolarization through the ventricles. Therefore, the P wave occurs immediately before the *beginning of contraction of the atria*, and the QRS wave occurs immediately before the *beginning of contraction of the ventricles*. The ventricles remain contracted until a few milliseconds after repolarization has occurred, that is, until after the end of the T wave

The atria repolarize approximately 0.15 to 0 20 second after the depolarization wave. However, this is just at the moment that the QRS wave is being recorded in the electrocardiogram Therefore, the atrial repolarization wave, known as the *atrial T wave*, is usually totally obscured by the much larger QRS wave. For this reason, an atrial T wave is rarely observed in the electrocardiogram.

On the other hand, the ventricular repolarization wave is the T wave of the normal electrocardiogram Ordinarily, ventricular muscle begins to repolarize in some fibers approximately 0 20 second after the beginning of the depolarization wave, but in many other fibers repolarization does not occur until as long as 0 35 second after onset of depolarization. Thus, the process of repolarization extends over a long period of time, about 0 15 second. For this reason the T wave in the normal electrocardiogram is often a prolonged wave, but the voltage of the T wave is considerably less than the voltage of the QRS complex, partly because of its prolonged length

VOLTAGE AND TIME CALIBRATION OF THE ELECTROCARDIOGRAM

All recordings of electrocardiograms are made with appropriate calibration lines on the recording paper. Either these calibration lines are already ruled on the paper, as is the case when a pen recorder is used, or they are recorded on the paper at the same time that the electrocardiogram is recorded, which is the case with the photographic types of electrocardiographs.

As illustrated in Figure 15–1, the horizontal calibration lines are arranged so that 10 small divisions upward or downward in the standard electrocardiogram represent 1 millivolt, with positivity in the upward direction and negativity in the downward direction.

The vertical lines on the electrocardiogram are time calibration lines Each inch in the horizontal direction

is 1 second, and each inch in turn is usually broken into five segments by dark vertical lines; the intervals between these lines represent 0.20 second. These intervals are then broken into five smaller intervals by thin lines, and each of these represents 0.04 second.

Normal Voltages in the Electrocardiogram. The voltages of the waves in the normal electrocardiogram depend on the manner in which the electrodes are applied to the surface of the body. When one electrode is placed directly over the heart and the second electrode is placed elsewhere on the body, the voltage of the QRS complex may be as great as 3 to 4 millivolts (mv). Even this voltage is very small in comparison with the monophasic action potential of 110 mv recorded directly at the heart muscle membrane. When electrocardiograms are recorded from electrodes on the two arms or on one arm and one leg, the voltage of the QRS complex usually is approximately 1 mv from the top of the R wave to the bottom of the S wave; the voltage of the P wave, between 0.1 and 0.3 mv; and that of the T wave, between 0.2 and 0.3 mv.

The P-Q or P-R Interval. The duration of time between the beginning of the P wave and the beginning of the QRS wave is the interval between the beginning of contraction of the atrium and the beginning of contraction of the ventricle. This period of time is called the P-Q interval. The normal P-Q interval is approximately 0.16 second. This interval is sometimes also called the P-R interval because the Q wave is frequently absent.

The Q-T Interval. Contraction of the ventricle lasts essentially from the beginning of the Q wave to the end of the T wave. This interval of time is called the Q-T interval and ordinarily is approximately 0.35 second.

The Rate of the Heart as Determined from Electrocardiograms. The rate of heartbeat can be determined easily from electrocardiograms, because the time interval between two successive beats is the reciprocal of the heart rate. If the interval between two beats as determined from the time calibration lines is 1 second, the heart rate is 60 beats per minute. The normal interval between two successive QRS complexes is approximately 0.83 second. This is a heart rate of 60/0.83 times per minute, or 72 beats per minute.

METHODS FOR RECORDING ELECTROCARDIOGRAMS

The electrical currents generated by the cardiac muscle during each beat of the heart sometimes change potentials and polarity in less than 0.01 second. Therefore, it is essential that any apparatus for recording electrocardiograms be capable of responding rapidly to these changes in electrical potentials. In general, two different types of recording apparatuses are used for this purpose, as follows:

THE PEN RECORDER

Most older types of electrocardiographic apparatus used optical and photographic methods for recording the electrocardiogram. In many of these a beam of light was focused on a mirror mounted on a galvanometer coil located in a powerful magnetic field. When current flowed through the coil, the coil rotated in the field, and the mirror swept the reflected beam of light back and forth across a moving photographic paper. However, because the records required photographic development before the recording could be viewed, direct pen writing recorders have become the vogue in recent years.

The pen writing recorders write the electrocardiogram with a pen directly on a moving sheet of paper. The pen is often a thin tube connected at one end to an inkwell, and its recording end is connected to a powerful electromagnet system that is capable of moving the pen back and forth at high speed. As the paper moves forward, the pen records the electrocardiogram. The movement of the pen in turn is controlled by means of appropriate amplifiers connected to electrocardiographic electrodes on the patient.

Other pen recording systems use special paper that does not require ink in the recording stylus. One such paper turns black when it is exposed to heat; the stylus itself is made very hot by electrical current flowing through its tip. Another type turns black when electrical current flows from the tip of the stylus through the paper to an electrode at its back. This leaves a black line at every point on the paper that the stylus touches.

RECORDING ELECTROCARDIOGRAMS WITH THE OSCILLOSCOPE

Electrocardiograms can also be viewed on the screen of an oscilloscope by the method discussed for nerve potentials in Chapter 10, or they can be photographed from the oscilloscopic screen. However, because of the cost of the oscilloscope, and because extremely high frequency electrical potentials do not need to be recorded, the less complicated and less expensive pen recorder just described is ordinarily used in clinical electrocardiography.

FLOW OF CURRENT AROUND THE HEART DURING THE CARDIAC CYCLE

RECORDING ELECTRICAL POTENTIALS FROM A PARTIALLY DEPOLARIZED MASS OF SYNCYTIAL CARDIAC MUSCLE

Figure 15-4 illustrates a syncytial mass of cardiac muscle that has been stimulated at its central most point.

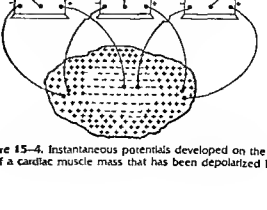


Figure 15-4. Instantaneous potentials developed on the surface of a cardiac muscle mass that has been depolarized in its center.

Prior to stimulation, all the exteriors of the muscle cells had been positive and the interiors, negative. However, for reasons presented in Chapter 10 in the discussion of membrane potentials, as soon as an area of the cardiac syncytium becomes depolarized, negative charges leak to the outsides of the depolarized muscle fibers, making this surface area electronegative, as represented by the negative signs in the figure, with respect to the remaining surface of the heart, which is still polarized in the normal manner, as represented by the positive signs. Therefore, a meter connected with its negative terminal on the area of depolarization and its positive terminal on one of the still-polarized areas, as illustrated to the right in the figure, will record positively.

Two other possible electrode placements and meter readings are also illustrated in Figure 15–4. *These should be studied carefully and the student should explain to himself or herself the causes of the respective meter readings.* Obviously, since the process of depolarization spreads in all directions through the heart, the potential differences shown in the figure last for only a few milliseconds, and the actual voltage measurements can be accomplished only with a high-speed recording apparatus.

FLOW OF ELECTRICAL CURRENTS AROUND THE HEART IN THE CHEST

Figure 15–5 illustrates the ventricular muscle mass lying within the chest. Even the lungs, though filled with air, conduct electricity to a surprising extent, and fluids of the other tissues surrounding the heart conduct electricity even more easily. Therefore, the heart is actually suspended in a conductive medium. When one portion of the ventricles becomes electronegative with respect to the remainder, electrical current flows from the depolarized area to the polarized area in large circuitous routes, as noted in the figure.

It will be recalled from the discussion of the Purkinje system in Chapter 14 that the cardiac impulse first arrives in the ventricles in the septum and shortly thereafter reaches the endocardial surfaces of the remainder of the ventricles, as shown by the colored areas and the negative signs in Figure 15–5. This provides electronegativity on the insides of the ventricles and electropositivity on the outer walls of the ventricles, and current flows through the fluids surrounding the ventricles along elliptical paths, as illustrated in the figure. If one algebraically averages all the lines of current flow (the elliptical lines), one finds that the average current flow is from the base of the heart toward the apex. During most of the remainder of the depolarization process, the current continues to flow in this direction as the depolarization spreads from the endocardial surface outward through the ventricular muscle. However, immediately before the depolarization has completed its course through the ventricles, the direction of current flow reverses for about 1/100 second, flowing then from the apex toward the base because the very last part of the heart to become depolarized is the outer walls of the ventricles near the base of the heart.

Thus, in the normal heart it may be considered that current flows primarily in the direction from the base toward the apex during almost the entire cycle of depolarization except at the very end. Therefore, if a meter is connected to the surface of the body as shown in Figure 15–5, the electrode nearer the base will be negative with respect to the electrode nearer the apex, and the recording meter will show a slight positive potential between the two electrodes. In making electrocardiographic recordings, various standard positions for placement of electrodes are used, and whether the polarity of the recording during each cardiac cycle is positive or negative is determined by the orientation of electrodes with respect to the current flow in the heart. Some of the conventional electrode systems, commonly called *electrocardiographic leads*, are discussed below.

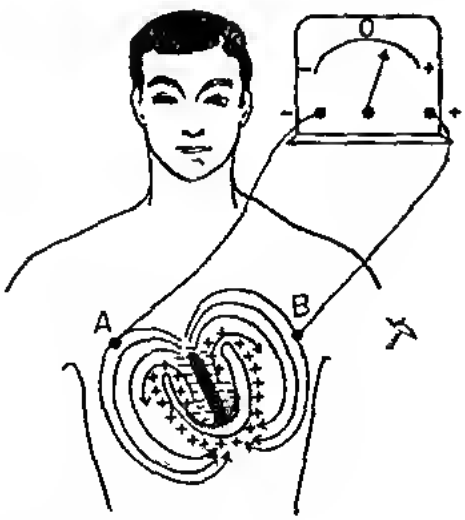


Figure 15–5. Flow of current in the chest around a partially depolarized heart.

ELECTROCARDIOGRAPHIC LEADS

THE THREE BIPOLAR LIMB LEADS

Figure 15–6 illustrates electrical connections between the limbs and the electrocardiograph for recording electrocardiograms from the so-called standard "bipolar" limb leads. The term "bipolar" means that the electrocardiogram is recorded from two specific electrodes on the body, in this case on the limbs. The electrocardiograph in each instance is illustrated by mechanical meters in the diagram, though the actual electrocardiograph is a high-speed recording meter.

Lead I. In recording limb lead I, the *negative terminal of the electrocardiograph is connected to the right arm* and the *positive terminal to the left arm*. Therefore, when the point on the chest where the right arm connects to the chest is electronegative with respect to the point where the left arm connects, the electrocardiograph records positively—that is, above the zero voltage line in the electrocardiogram. When the opposite is true, the electrocardiograph records below the line.

Lead II. In recording limb lead II, the *negative terminal of the electrocardiograph is connected to the right arm* and the *positive terminal to the left leg*. Therefore, when the right arm is negative with respect to the left leg, the electrocardiograph records positively.

Lead III. In recording limb lead III, the *negative terminal of the electrocardiograph is connected to the left arm* and the *positive terminal to the left leg*. This means

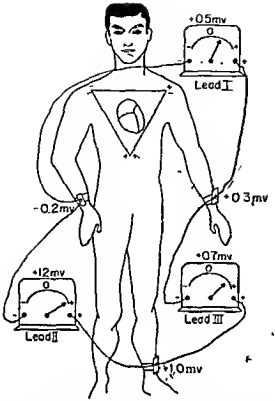


Figure 15–6. Conventional arrangement of electrodes for recording the standard electrocardiographic leads. Einthoven's triangle is superimposed on the chest.

that the electrocardiograph records positively when the left arm is negative with respect to the left leg.

Einthoven's Triangle. In Figure 15–6, a triangle, called *Einthoven's triangle*, is drawn around the area of the heart. This is a diagrammatic means of illustrating that the two arms and the left leg form apices of a triangle surrounding the heart. The two apices at the upper part of the triangle represent the points at which the two arms connect electrically with the fluids around the heart, and the lower apex is the point at which the left leg connects with the fluids.

Einthoven's Law. Einthoven's law states simply that if the electrical potentials of any two of the three bipolar limb electrocardiographic leads are known at any given instant, the third one can be determined mathematically from the first two by simply summing the first two (but note that the positive and negative signs of the different leads must be observed when making this summation). For instance, let us assume that momentarily, as noted in Figure 15–6, the right arm is 0.2 mv negative with respect to the average potential in the body, the left arm is 0.3 mv positive, and the left leg is 1.0 mv positive. Observing the meters in the figure, it will be seen that lead I records a positive potential of 0.5 mv because this is the difference between the −0.2 mv on the right arm and the +0.3 mv on the left arm. Similarly, lead III records a positive potential of 0.7 mv and lead II records a positive potential of 1.2 mv because these are the instantaneous potential differences between the respective pairs of limbs.

Note that *the sum of the voltages in leads I and III equals the voltage in lead II.* That is, 0.5 plus 0.7 equals 1.2. Mathematically, this principle, called Einthoven's law, holds true at any given instant while the electrocardiogram is being recorded.

Normal Electrocardiograms Recorded from the Three Bipolar Limb Leads. Figure 15–7 illustrates simultaneous recordings of the electrocardiogram in leads I, II, and III. It is obvious from this figure that the electrocardiograms in these three leads are very similar to each other, for they all record positive *P* waves and positive *T* waves, and the major portion of the QRS complex is also positive in each electrocardiogram.

On analysis of the three electrocardiograms, it can be shown with careful measurements that at any given instant the sum of the potentials in leads I and III equals the potential in lead II, thus illustrating the validity of Einthoven's law.

Because the recordings from all the bipolar limb leads are similar to each other, it does not matter greatly which lead is recorded when one wishes to diagnose the different cardiac arrhythmias, for diagnosis of arrhythmias depends mainly on the time relationships between the different waves of the cardiac cycle. On the other hand, when one wishes to diagnose damage in the ventricular or atrial muscle or in the conducting system, it does matter greatly which leads are recorded, for abnormalities of the cardiac muscle change the patterns of the electrocardiograms markedly in some leads and yet may not affect other leads.

Electrocardiographic interpretation of these two types of conditions—cardiac myopathies and cardiac arrhythmias—is discussed separately in the following two chapters.

CHEST LEADS (PRECORDIAL LEADS)

Often electrocardiograms are recorded with an electrode placed on the anterior surface of the chest over the heart at one of the six separate red-colored points in Figure 15–8. This electrode is connected to the positive terminal of the electrocardiograph, while the negative electrode, called the *indifferent electrode*, is normally connected simultaneously through electrical resistances to the right arm, left arm, and left leg, as also shown in the figure. Usually six different standard chest leads are recorded from the anterior chest wall, the chest electrode being placed respectively at the six

Figure 15–7. Normal electrocardiograms recorded from the three standard electrocardiographic leads.

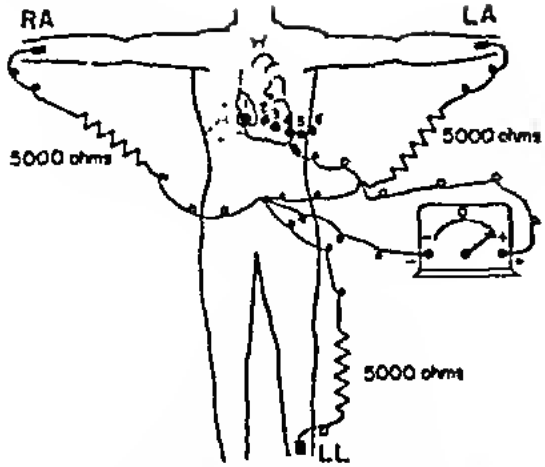


Figure 15–8. Connections of the body with the electrocardiograph for recording chest leads

points illustrated in the diagram. The different recordings recorded by the method illustrated in Figure 15–8 are known as leads V_1, V_2, V_3, V_4, V_5, and V_6.

Figure 15–9 illustrates the electrocardiograms of the normal heart as recorded from these six standard chest leads. Because the heart surfaces are close to the chest wall, each chest lead records mainly the electrical potential of the cardiac musculature immediately beneath the electrode. Therefore, relatively minute abnormalities in the ventricles, particularly in the anterior ventricular wall, frequently cause marked changes in the electrocardiograms recorded from chest leads.

In leads V_1 and V_2, the QRS recordings of the normal heart are mainly negative because, as illustrated in Figure 15–8, the chest electrode in these leads is nearer the base of the heart than the apex, which is the direction of electronegativity during most of the ventricular depolarization process. On the other hand, the

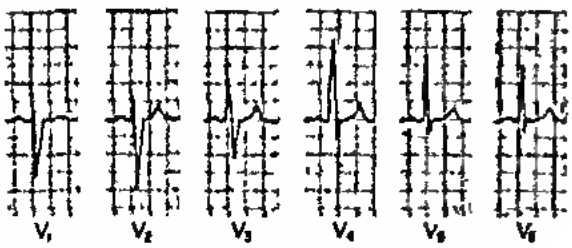


Figure 15–9. Normal electrocardiograms recorded from the six standard chest leads.

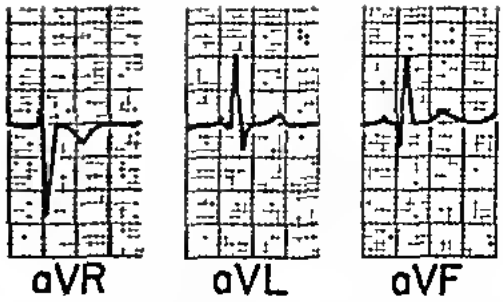


Figure 15–10. Normal electrocardiograms recorded from the three augmented unipolar limb leads

QRS complexes in leads V_4, V_5, and V_6 are mainly positive because the chest electrode in these leads is nearer the apex, which is the direction of electropositivity during most of depolarization.

AUGMENTED UNIPOLAR LIMB LEADS

Another system of leads in wide use is the "augmented unipolar limb lead." In this type of recording, two of the limbs are connected through electrical resistances to the negative terminal of the electrocardiograph while the third limb is connected to the positive terminal. When the positive terminal is on the right arm, the lead is known as the aV_R lead, when on the left arm, the aV_L lead, and when on the left leg, the aV_F lead.

Normal recordings of the augmented unipolar limb leads are shown in Figure 15–10. These are all similar to the standard limb lead recordings except that the recording from the aV_R lead is inverted. The reason for this inversion is that the polarity of the electrocardiograph in this instance is connected backward to the major direction of current flow in the heart during the cardiac cycle.

Each augmented unipolar limb lead records the potential of the heart on the side nearest to the respective limb. Thus, when the recording in the aV_R lead is negative, this means that the side of the heart nearest to the right arm is negative in relation to the remainder of the heart; when the recording in the aV_F lead is positive, this means that the apex of the heart, which is the part of the heart nearest the left foot, is positive with respect to the remainder of the heart.

REFERENCES

See bibliography at end of Chapter 17

16

Electrocardiographic Interpretation in Cardiac Muscle Abnormalities—Vectorial Analysis

From the discussion in Chapter 14 of impulse transmission through the heart, it is obvious that any change in the pattern of this transmission can cause abnormal electrical currents around the heart and, consequently, can alter the shapes of the waves in the electrocardiogram. For this reason, almost all serious abnormalities of the heart muscle can be detected by analyzing the contours of the different waves in the different electrocardiographic leads. The purpose of the present chapter is to discuss the various alterations in the electrocardiograms when either the muscle of the heart or the conduction system, especially of the ventricles, functions abnormally.

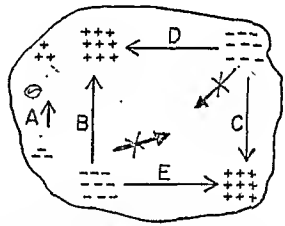


Figure 16–1. Vectors depicting current flow within a mass of cardiac muscle. See text for discussion.

PRINCIPLES OF VECTORIAL ANALYSIS OF ELECTROCARDIOGRAMS

USE OF VECTORS TO REPRESENT ELECTRICAL POTENTIALS

Before it is possible to understand how cardiac abnormalities affect the contours of waves in the electrocardiogram, one must first become familiar with the concept of vectors and vectorial analysis as applied to electrical currents flowing in and around the heart.

Several times in the preceding chapter it was pointed out that heart currents flow in a particular direction at a given instant in the cardiac cycle. A vector is an arrow that points in the direction of current flow *with the arrowpoint in the positive direction*. Also, by convention, the length of the arrow is drawn *proportional to the voltage generated by the current flow*.

In Figure 16–1, several different vectors are shown between areas of negativity and positivity in a syncytial mass of cardiac muscle. Note, first, that each vector points from negative to positive and, second, that the length of each vector is proportional to the amount of charges causing the current flow.

The "Resultant" Vector in the Heart at Any Given Instant. Figure 16–2 shows, via the shaded area and the negative signs, depolarization of the ventricular septum and parts of the lateral endocardial walls of the two ventricles. Electrical currents flow between these depolarized areas inside the heart and the nondepolarized areas on the outside of the heart as indicated by the elliptical arrows. Currents also flow inside the heart chambers directly from the depolarized areas toward the polarized areas. Even though a small amount of current flows upward inside the heart, a considerably greater quantity flows downward on the outside of the ventricles toward the apex. Therefore, the summated vector of currents at this particular instant, called the *instantaneous "resultant" vector*, is drawn through the center of the ventricles in a direction from the base of the heart toward the apex. Furthermore, because these currents are considerable in quantity, the vector is relatively long.

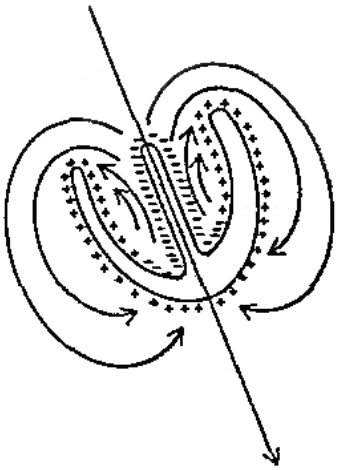


Figure 16–2. A resultant vector through the partially depolarized heart

DENOTING THE DIRECTION OF A VECTOR IN TERMS OF DEGREES

When a vector is horizontal and directed toward the subject's left side, it is said that the vector extends in the direction of 0 degrees, as is illustrated in Figure 16–3. From this zero reference point, the scale of vectors rotates clockwise, when the vector extends from above downward, it has a direction of 90 degrees; when it extends from the subject's left to the right, it has a direction of 180 degrees; and, when it extends upward, it has a direction of −90 or +270 degrees.

In a normal heart the average direction of the vector of the heart during spread of the depolarization wave through the ventricles, called the *mean QRS vector*, is approximately 59 degrees, which is illustrated by vector *A* drawn through the center of Figure 16–3 in the 59 degree direction. This means that during most of the depolarization wave, the apex of the heart remains positive with respect to the base of the heart, as is discussed later in the chapter.

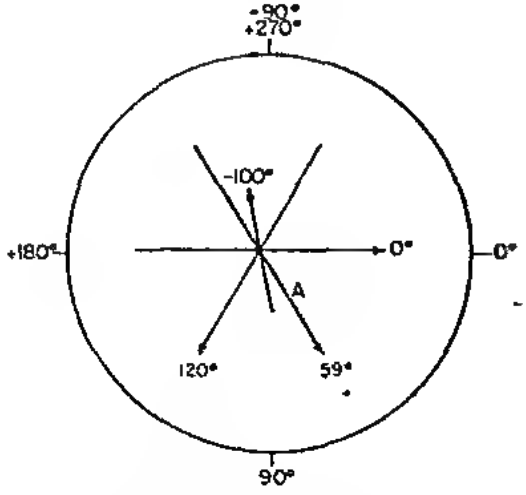


Figure 16–3. Vectors drawn to represent direction of current flow and potentials for several different hearts.

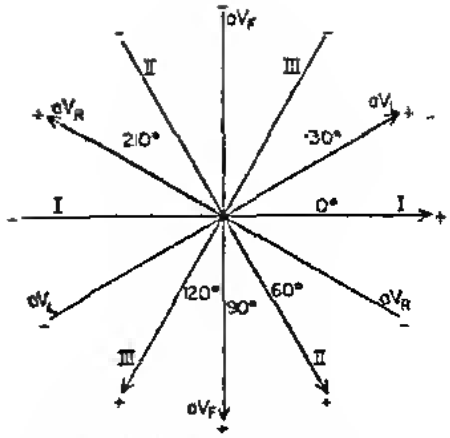


Figure 16–4. Axes of the three bipolar and three unipolar leads.

"AXIS" OF EACH OF THE BIPOLAR AND UNIPOLAR LIMB LEADS

In the preceding chapter the three bipolar and three unipolar limb leads were described. Each lead is actually a pair of electrodes connected to the body on opposite sides of the heart, and the direction from the negative to the positive electrode is called the *axis of the lead*. Lead I is recorded from two electrodes placed respectively on the two arms. Since the electrodes lie in the horizontal direction with the positive electrode to the left, the axis of lead I is 0 degrees.

In recording lead II, electrodes are placed on the right arm and left leg. The right arm connects to the torso in the upper right-hand corner and the left leg to the lower left-hand corner. Therefore, the direction of this lead is approximately 60 degrees.

By a similar analysis it can be seen that lead III has an axis of approximately 120 degrees, lead aV_R 210 degrees, aV_F 90 degrees, and aV_L −30 degrees. The directions of the axes of all these different leads are shown in the diagram of Figure 16–4, which is known as the *hexagonal reference system*. The polarities of the electrodes are illustrated by the plus and minus signs. *The student must learn these axes and their polarities, particularly for the bipolar limb leads I, II, and III, in order to understand the remainder of this chapter.*

VECTORIAL ANALYSIS OF POTENTIALS RECORDED IN DIFFERENT LEADS

Now that we have discussed, first, the conventions for representing current flow and potentials across the heart by means of vectors and, second, the axes of the leads, it becomes possible to put these two together to determine the potential that will be recorded in each lead for a given vector in the heart.

Figure 16–5 illustrates a partially depolarized heart; vector *A* represents the instantaneous mean direction of current flow in the heart and its potential. In this instance the direction of current flow is 55 degrees, and the potential will be assumed to be 2 mv. Through the base of vector *A* is drawn the axis of lead I in the 0 degree direction. From the tip of vector *A* a perpendicular is dropped to the lead I axis, and a so-called *projected vector (B)* is drawn along the axis. The head of this vector is in the positive direction, which means that the record momentarily being recorded in the elec-

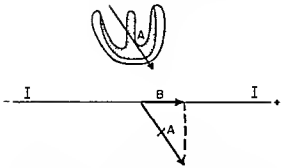


Figure 16–5. Determination of a projected vector B along the axis of lead I when vector A represents the instantaneous current flow in the ventricles.

trocardiogram of lead I will be positive. The voltage recorded will be equal to the length of B divided by the length of A times 2 mv, or approximately 1 mv.

Figure 16–6 illustrates another example of vectorial analysis. In this example, vector A represents the current flow at a given instant during ventricular depolarization in another heart, in which the left side of the heart becomes depolarized somewhat more rapidly than the right. In this instance the vector has a direction of 100 degrees, and the voltage is again 2 mv. To determine the potential in lead I, we drop a perpendicular to the lead I axis and find the projected vector B. Vector B is very short and this time in the negative direction, indicating that at this particular instant the recording in lead I will be negative (below the zero line), and the voltage recorded will be slight. This figure illustrates that *when the vector in the heart is in a direction almost perpendicular to the axis of the lead, the voltage recorded in the electrocardiogram of this lead is very low.* On the other hand, *when the heart vector has almost exactly the same axis as the lead, essentially the entire voltage of the vector will be recorded.*

Vectorial Analysis of Potentials in the Three Standard Bipolar Limb Leads. In Figure 16–7, vector A depicts the instantaneous direction and potential of current in a partially depolarized ventricle. To determine the potential recorded at this instant in the electrocardiogram of each one of the three standard bipolar limb leads, perpendiculars are dropped to all the lines representing the different leads as illustrated in the figure. The projected vector B depicts the potential

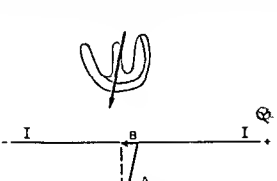


Figure 16–6. Determination of the projected vector B along the axis of lead I when vector A represents the instantaneous current flow in the ventricles.

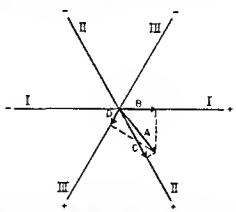


Figure 16–7. Determination of projected vectors in leads I, II, and III when vector A represents the instantaneous current flow in the ventricles.

recorded at that instant in lead I; projected vector C depicts the potential in lead II; and projected vector D, in lead III. In each of these the record in the electrocardiogram is positive—that is, above the zero line—because the projected vectors lie in the positive directions along the axes of all the leads. The potential in lead I is approximately one-half that of the vector through the heart, in lead II it is almost exactly equal to that in the heart, and in lead III it is about one-third that in the heart.

An identical analysis can be used to determine potentials recorded in augmented limb leads, except that the respective axes of these leads (see Figure 16–4) are used in place of the standard bipolar limb lead axes used in Figure 16–7.

VECTORIAL ANALYSIS OF THE NORMAL ELECTROCARDIOGRAM

VECTORS OCCURRING DURING DEPOLARIZATION OF THE VENTRICLES—THE QRS COMPLEX

When the cardiac impulse enters the ventricles through the A-V bundle, the first part of the ventricles to become depolarized is the left endocardial surface of the septum. This depolarization spreads rapidly to involve both endocardial surfaces of the septum, as illustrated by the shaded portion of the ventricle in Figure 16–8A. Then the depolarization spreads along the endocardial surfaces of the two ventricles, as shown in Figures 16–8B and C. Finally, it spreads through the ventricular muscle to the outside of the heart, as shown progressively in Figures 16–8C, D; and E.

At each stage of depolarization of the ventricles in Figures 16–8A to E, the instantaneous, summated current flow is represented by a vector superimposed on the ventricle in each figure. Each of these vectors is analyzed by the method described in the preceding section to determine the voltages that will be recorded at each instant in each of the three standard electrocardiographic leads. To the right in each stage of the figure is shown the progressive development of the QRS complex. *Keep in mind that a positive vector in a lead will cause the recording in the electrocardiogram to be above*

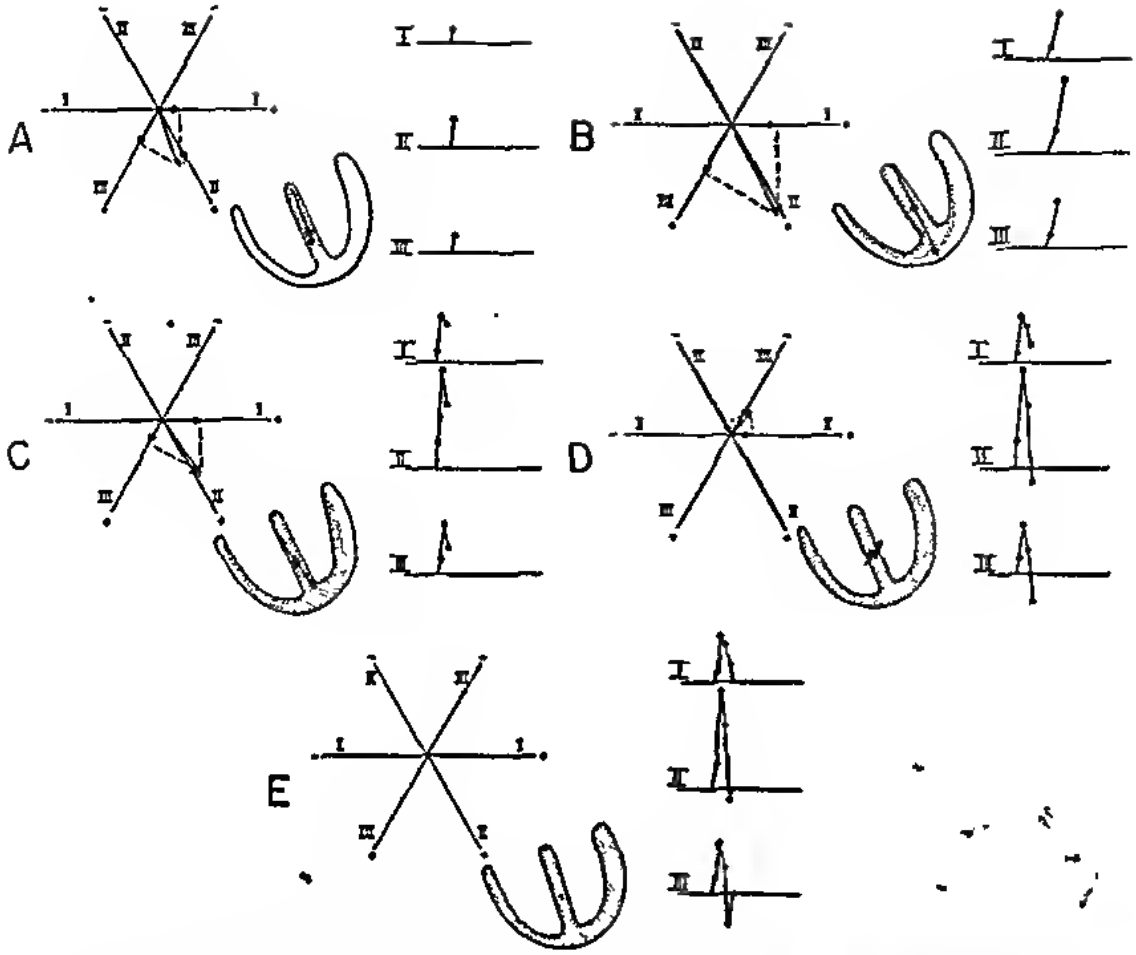


Figure 16–8. *A*, The ventricular vectors and the QRS complexes 0.01 second after onset of ventricular depolarization. *B*, 0.02 second after onset of depolarization. *C*, 0.035 second after onset of depolarization. *D*, 0.05 second after onset of depolarization and *E* after depolarization of the ventricles is complete, 0.06 second after onset.

the zero line, while a negative vector will cause the recording to be below the line.

Before proceeding with any further consideration of vectorial analysis, it is essential that this analysis of the successive normal vectors presented in Figure 16–8 be understood. Each of these analyses should be studied in detail by the procedure given above. A short summary of this sequence is the following.

In Figure 16–8A the ventricular muscle has just begun to be depolarized, representing an instant about 0.01 second after the onset of depolarization. At this time, the vector is short because only a small portion of the ventricles—the septum—is depolarized. Therefore, all electrocardiographic voltages are low, as recorded to the right of the ventricular muscle for each of the leads. The voltage in lead II is greater than the voltages in leads I and III because the heart vector extends mainly in the same direction as the axis of lead II.

In Figure 16–8B, which represents approximately 0.02 second after onset of depolarization, the heart vector is [illegible] because much of the ventricles has now become [illegible]ized. Therefore, the voltages in all electrocardi[illegible] leads have increased.

In Figure 16–8C, about 0.035 second after onset of [illegible]larization, the heart vector is becoming shorter and the recorded EKG voltages are less because the outside of the apex of the heart is now electronegative, neutralizing much of the negativity on the endocardial surfaces of the heart. Also, the axis of the vector is shifting toward the left side of the chest because the left ventricle is slightly slower to depolarize than the right. The ratio of the voltage in lead I to that in lead III is increasing.

In Figure 16–8D, about 0.05 second after onset of depolarization, the heart vector points toward the base of the left ventricle, and it is short because only a minute portion of the ventricular muscle is still polarized. Because of the direction of the vector at this time, the voltages recorded in leads II and III are both negative—that is, below the line.

In Figure 16–8E, about 0.06 second after onset of depolarization, the entire ventricular muscle mass is depolarized so that no current flows around the heart at all. The vector becomes zero, and the voltages in all leads become zero.

Thus, the QRS complexes are completed in the three standard bipolar limb leads. The QRS complex sometimes has a slight negative depression at its beginning in one or more of the leads, which is not shown in Figure 16–8; this is the Q wave. When it occurs, it is caused by initial depolarization of the left side of the septum before the right side, which creates a weak vector from left to right for a fraction of a second before the usual base-to-

apex vector occurs. The major positive deflection shown in Figure 16–8 is the R wave, and the final negative deflection is the S wave.

THE ELECTROCARDIOGRAM DURING REPOLARIZATION—THE T WAVE

Once the ventricular muscle has become depolarized, approximately 0.15 second elapses before sufficient repolarization begins for it to be observed in the electrocardiogram; then repolarization proceeds throughout the ventricular muscle until it is complete at about 0.35 second after the onset of the QRS complex. It is this process of repolarization that causes the T wave in the electrocardiogram.

Because the septum and other endocardial areas of the ventricular muscle depolarize first, it seems logical that these areas should repolarize first as well, but this is not the usual case because the septum and the other endocardial areas have a longer period of contraction and depolarization than do other areas of the heart. Actually, *the greatest portion of ventricular muscle to repolarize first is that located over the entire outer surface of the ventricles and especially near the apex of the heart.* The endocardial areas, on the other hand, normally repolarize last. The reason for this abnormal sequence of repolarization is believed to be that high pressure in the ventricles during contraction greatly reduces coronary blood flow to the endocardium, thereby slowing the repolarization process in the endocardial areas.

Therefore, because the outer and apical surfaces of the ventricles repolarize before the inner and basal surfaces, the positive end of the heart vector during repolarization is toward the apex of the heart. Thus, *the predominant direction of the vector through the heart during repolarization of the ventricles is from base to apex, which is also the predominant direction of the vector during depolarization. As a result, the T wave in the normal bipolar limb leads is positive, which is also the polarity of most of the normal QRS complex.*

In Figure 16–9 five stages of repolarization of the ventricles are denoted by the progressive increase of the white areas—the repolarized areas. At each stage, the vector extends from the base toward the apex until it disappears in the last stage. At first the vector is relatively small because the area of repolarization is small. Later the vector becomes stronger and stronger because of greater degrees of repolarization. Finally, the vector becomes weaker again because the areas of depolarization still persisting become so slight that the total quantity of current flow begins to decrease. These changes illustrate that the vector is greatest when approximately half the heart is in the polarized state and approximately half is depolarized.

The changes in the electrocardiograms of the three standard limb leads during the process of repolarization are noted under each of the ventricles, depicting the progressive stages of repolarization. Over approximately 0.15 second, the period required for the whole process to take place, the T wave of the electrocardiogram is generated.

DEPOLARIZATION OF THE ATRIA—THE P WAVE

Depolarization of the atria begins in the S-A node and spreads in all directions over the atria. Therefore, the point of original electronegativity in the atria is approximately at the base of the superior vena cava where the S-A node lies, and the direction of current flow in the atrium at the beginning of depolarization is in the direction noted in Figure 16–10. Furthermore, the vector remains generally in this direction throughout the process of atrial depolarization.

Thus, the vector of current flow during depolarization in the atria points in almost the same direction as that in the ventricles. And, because this direction is in the directions of the axes of the standard bipolar limb leads I, II, and III, the electrocardiograms recorded from the atria during the process of depolarization usually are positive in all three of these leads, as illustrated in Figure 16–10. The record of atrial depolarization is known as the P wave.

Repolarization of the Atria—The Atrial T Wave. Spread of the depolarization wave through the atrial muscle is *much slower than in the ventricles.* Therefore,

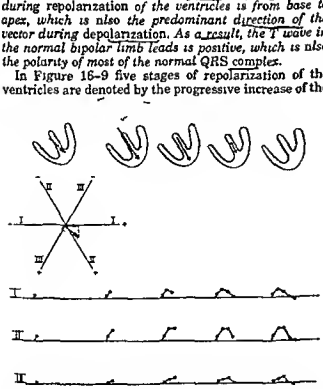


Figure 16–9. Generation of the T wave during repolarization of the ventricles, showing a vectorial analysis of the first stage. The total time from the beginning of the T wave to the end is about 0.15 second.

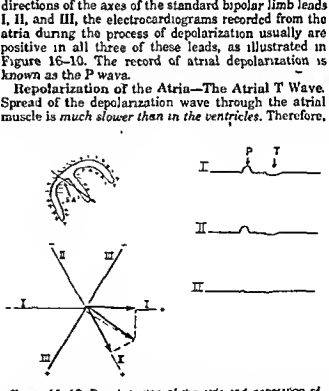


Figure 16–10. Depolarization of the atria and generation of the P wave, showing the vector through the atria and the resultant vectors in the three standard leads. At right are the atrial P and T waves.

the musculature around the S-A node becomes depolarized a long time before the musculature in distal parts of the atria. Because of this, *the area in the atria that becomes repolarized first is the S-A nodal region, the area that had originally become depolarized first*, which is an entirely different situation from that in the ventricles. Thus, when repolarization begins, the region around the S-A node becomes positive with respect to the rest of the atria. Therefore, the atrial repolarization vector is *backward to the vector of depolarization*. (Note again that this is opposite to the effect that occurs in the ventricles.) Therefore, as noted to the right in Figure 16–10, the so-called atrial T wave follows about 0.15 second after the atrial P wave, but this T wave is on the opposite side of the zero reference line from the P wave—that is, it is normally negative rather than positive in the three standard bipolar limb leads. In the normal electrocardiogram this T wave appears at approximately the same time that the QRS complex of the ventricle appears. Therefore, it is almost always totally obscured by the larger QRS complex, though in some abnormal states it does play a role in the recorded electrocardiogram.

THE VECTORCARDIOGRAM

It was noted in the preceding discussions that the vector of current flow through the heart changes rapidly as the impulse spreads through the myocardium. It changes in two aspects. First, the vector increases and decreases in length because of the increasing and decreasing potential of the vector. Second, the vector changes its direction because of changes in the average direction of current flow around the heart. The so called *vectorcardiogram* depicts these changes in the vectors at the different times during the cardiac cycle, as illustrated in Figure 16–11.

In the vectorcardiogram of Figure 16–11, point 5 is the *zero reference point*, and this point is the negative end of all the vectors. While the heart is quiescent, the positive end of the vector also remains at the zero point because there is no current flow. However, just as soon as current begins to flow through the heart, the positive end of the vector leaves the zero reference point.

When the septum first becomes depolarized, the vector extends downward toward the apex of the heart, but it is relatively weak, thus generating the first portion of the vectorcardiogram, as illustrated by the positive end of vector 1. As more of the heart becomes depolarized, the vector becomes stronger and stronger, usually swinging slightly to one side. Thus vector 2 of Figure 16–11 represents the state of depolarization of the heart about 0.02 second after vector 1. After another 0.02 second, vector 3 represents the current flow in the heart, and vector 4 occurs in still another 0.01 second. Finally, the heart becomes totally depolarized, and the vector becomes zero once again, as shown at point 5.

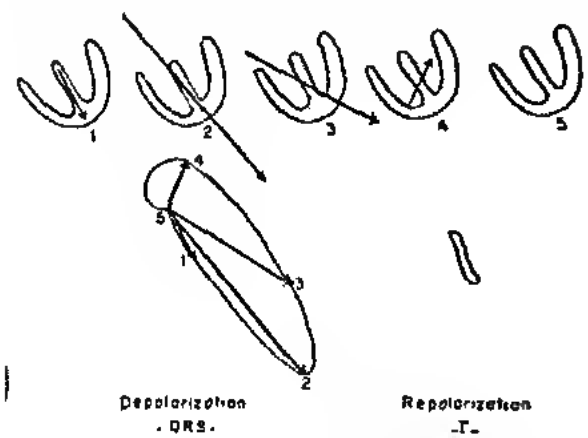


Figure 16–11. The QRS and T vectorcardiograms.

The elliptical figure generated by the positive ends of the vectors is called the *QRS vectorcardiogram*.

Vectorcardiograms can be recorded instantaneously on an oscilloscope by connecting electrodes from above and below the heart to the vertical plates of the oscilloscope and connecting electrodes from each side of the heart to the horizontal plates. When the vector changes, the spot of light on the oscilloscope follows the rourse of the positive end of the changing vector, thus inscribing the vectorcardiogram on the oscilloscopic screen.

The "T" Vectorcardiogram. Changing vectors in the heart occur not only during the depolarization process, for vectors depicting current flow around the ventricles also reappear during repolarization. Therefore, a second, smaller vectorcardiogram—the *T vectorcardiogram*—is inscribed during repolarization of the muscular mass; this is depicted to the right in Figure 16–10. Also, a still smaller *P vectorcardiogram* is inscribed during atrial depolarization.

THE MEAN ELECTRICAL AXIS OF THE VENTRICLE

The vectorcardiogram of the ventricular depolarization wave (the QRS vectorcardiogram) shown in Figure 16–11 is that of a normal heart. Note from this vectorcardiogram that the preponderant direction of the vectors of the ventricles is normally toward the apex of the heart—that is, during most of the cycle of ventricular depolarization, current flows from the base of the ventricles toward the apex. This preponderant direction of current flow during depolarization is called the *mean electrical axis of the ventricles* or the *mean QRS vector*. The mean electrical axis of the normal ventricles is 59 degrees. However, in certain pathological conditions of the heart this direction of current flow is changed markedly—sometimes even to opposite poles of the heart.

DETERMINING THE ELECTRICAL AXIS FROM STANDARD LEAD ELECTROCARDIOGRAMS

Clinically, the electrical axis of the heart is usually determined from the standard bipolar limb lead electrocardiograms rather than from the vectorcardiogram. Figure 16–12 illustrates a method for doing this. After recording the various standard leads, one determines the maximum potential and polarity of the recording in two of the leads. In lead I of the figure the recording is positive and in lead III the recording is mainly positive, but negative during part of the cycle. If any part of the recording is negative, *this negative potential is subtracted from the positive potential* to determine the net potential for that lead, as illustrated by the arrows to the right of the QRS complexes of leads I and III. (To be even more accurate, most cardiologists subtract the *area* of the negative wave from the *area* of the positive wave.)

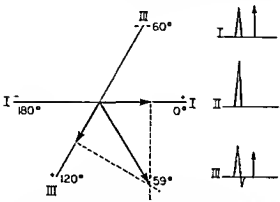


Figure 16–12. Plotting the mean electrical axis of the heart from two electrocardiographic leads.

After subtracting the negative portion of the QRS wave in lead III from the positive portion, each net potential is plotted on the axes of the respective leads, with the base of the potential at the point of intersection of the axes, as illustrated in Figure 16–12.

If the net potential of lead I is positive, it is plotted in a positive direction along the line depicting lead I. On the other hand, if this potential is negative, it is plotted in a negative direction. Also, for lead III, the net potential is placed with its base at the point of intersection, and, if positive, it is plotted in the positive direction along the line depicting lead III. If it is negative it is plotted in the negative direction.

To determine the actual vector of current flow in the ventricles, one draws perpendicular lines from the apices of the two net potentials of leads I and III, respectively. The point of intersection of these two perpendicular lines represents, by vectorial analysis, the apex of the actual mean QRS vector in the ventricles, and the point of intersection of the two lead axes represents the negative end of the actual vector. Therefore, the *mean QRS vector* is drawn between these two points. The average potential generated by the ventricles during depolarization is represented by the length of the vector, and the mean electrical axis is represented by the direction of the vector. Thus, the orientation of the mean electrical axis of the normal ventricles, as determined in Figure 16–12, is 59 degrees.

ABNORMAL VENTRICULAR CONDITIONS THAT CAUSE AXIS DEVIATION

Though the mean electrical axis of the ventricles averages approximately 59 degrees, this axis can swing to the left even in the normal heart to approximately 20 degrees or to the right to approximately 100 degrees. The causes of the normal variations are simply anatomical differences in the Purkinje system distribution or in the musculature itself of different hearts. Yet, a number of conditions can cause axis deviation even beyond these normal limits, as follows:

Change in the Position of the Heart. Obviously, if the heart itself is angulated to the left, the mean electrical axis of the heart will also shift to the left. Such shift occurs (a) during expiration, (b) when a person lies down, because the abdominal contents press upward against the diaphragm, and (c) quite frequently in stocky, fat persons whose diaphragms normally press upward against the heart all the time.

Likewise, angulation of the heart to the right causes the mean electrical axis of the ventricles to shift to the right. This condition occurs (a) during inspiration, (b) when a person stands up, and (c) normally in tall, lanky persons whose hearts hang downward.

Hypertrophy of One Ventricle. When one ventricle greatly hypertrophies, the axis of the heart shifts toward the hypertrophied ventricle for two reasons: First, far greater quantity of muscle exists on the hypertrophied side of the heart than on the other side, and this allows excess generation of electrical currents on that side. Second, more time is required for the depolarization wave to travel through the hypertrophied ventricle than through the normal ventricle. Consequently, the normal ventricle becomes depolarized—that is, becomes negative—considerably in advance of the hypertrophied ventricle, and this causes a strong vector from the normal side of the heart toward the hypertrophied side, which is still positively charged. Thus the axis deviates toward the hypertrophied ventricle.

Vectorial Analysis of Left Axis Deviation Resulting from Hypertrophy of the Left Ventricle. Figure 16–13 illustrates the three standard bipolar limb leads of an electrocardiogram in which an analysis of the axis direction shows left axis deviation with the mean electrical axis pointing in the −15 degree direction. This is a typical electrocardiogram resulting from increased muscular mass of the left ventricle. In this instance the axis deviation was caused by *hypertension* (high blood pressure), which caused the left ventricle to hypertrophy in order to pump blood against the elevated systemic arterial pressure. However, a similar picture of left axis deviation occurs when the left ventricle hypertrophies as a result of *aortic valvular stenosis, aortic valvular regurgitation,* or any of a number of *congenital heart conditions* in which the left ventricle enlarges while the right side of the heart remains relatively normal in size.

Vectorial Analysis of Right Axis Deviation Resulting from Hypertrophy of the Right Ventricle.

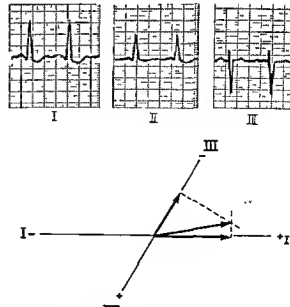


Figure 16–13. Left axis deviation in hypertensive heart disease. Note the slightly prolonged QRS complex as well.

The electrocardiogram of Figure 16–14 illustrates intense right axis deviation with an electrical axis of approximately 170 degrees, which is 111 degrees to the right of the normal mean electrical axis of the ventricles. The right axis deviation illustrated in this figure was caused by hypertrophy of the right ventricle as a result of *pulmonary stenosis*. However, right axis deviation may also occur in other congenital heart conditions that cause hypertrophy of the right ventricle, such as *tetralogy of Fallot* or *interventricular septal defect*. Also, hypertrophy of the right ventricle as a result of *increased pulmonary vascular resistance* can cause right axis deviation.

Bundle Branch Block. Ordinarily, the two lateral walls of the ventricles depolarize at almost the same time, because both the left and right bundle branches of the Purkinje system transmit the cardiac impulse to the endocardial surfaces of the two ventricular walls at almost the same instant. As a result, the currents flowing from the walls of the two ventricles almost neutralize each other. However, if one of the major bundle branches is blocked, the cardiac impulse spreads through one ventricle long before it spreads through the other. Therefore, depolarization of the two ventricles does not occur even nearly simultaneously, and the depolarization currents do not neutralize each other. As a result, axis deviation occurs as follows.

Vectorial Analysis of Left Bundle Branch Block. When the left bundle branch is blocked, cardiac depolarization spreads through the right ventricle two to three times as rapidly as through the left ventricle. Consequently, much of the left ventricle remains polarized for a long time after the right ventricle has become totally depolarized. Thus, the right ventricle becomes electronegative while the left ventricle remains electropositive during most of the depolarization process, and a very strong vector projects from the right ventricle toward the left ventricle. In other words, there is intense left axis deviation because the positive end of the vector points toward the left ventricle. This is illustrated in Figure 16–15, which shows typical left axis deviation resulting from left bundle branch block. Note that the axis is approximately −50 degrees.

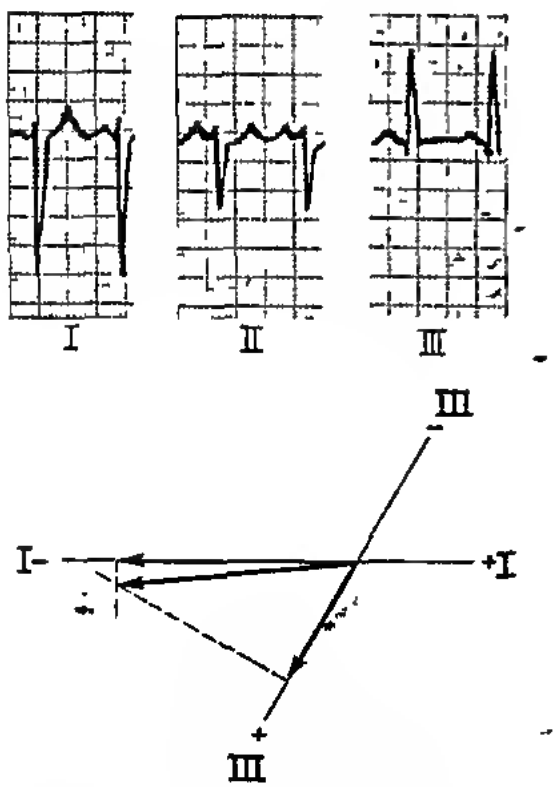


Figure 16–14. High voltage electrocardiogram in pulmonary stenosis with right ventricular hypertrophy. Intense right axis deviation is also seen, and a slightly prolonged QRS complex as well.

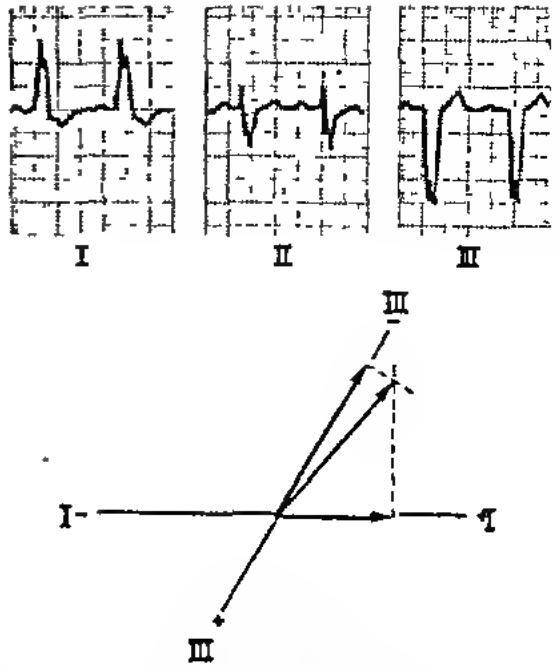


Figure 16–15. Left axis deviation due to left bundle branch block. Note the greatly prolonged QRS complex.

Because of slowness of impulse conduction when the Purkinje system is blocked, axis deviation resulting from bundle branch block also greatly increases the duration of the QRS complex, which one can see by observing the excessive widths of the QRS waves in Figure 16–15. This will be discussed in greater detail later in the chapter. This prolonged QRS complex differentiates this condition from axis deviation caused by hypertrophy.

Vectorial Analysis of Right Bundle Branch Block. When the right bundle branch is blocked, the left ven-

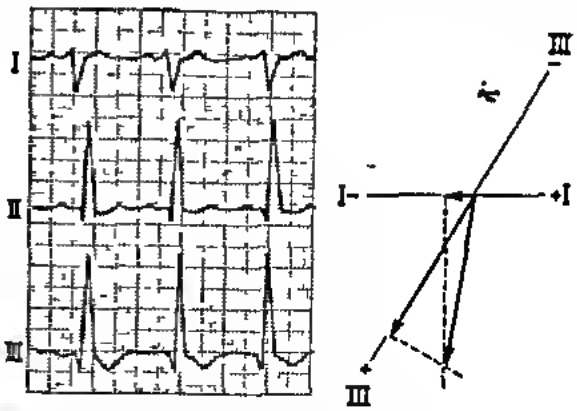


Figure 16–16. Right axis deviation due to right bundle branch block. Note the greatly prolonged QRS complex.

tricle depolarizes far more rapidly than the right ventricle so that the left becomes electronegative while the right remains electropositive. Therefore, a very strong vector develops with its negative end toward the left ventricle and its positive end toward the right ventricle. In other words, intense right axis deviation occurs.

Right axis deviation caused by right bundle branch block is illustrated and its vector analyzed in Figure 16–16, which shows an axis of approximately 105 degrees and a prolonged QRS complex because of blocked conduction. Note the great prolongation of the QRS complex.

CONDITIONS THAT CAUSE ABNORMAL VOLTAGES OF THE QRS COMPLEX

INCREASED VOLTAGE IN THE STANDARD BIPOLAR LIMB LEADS

Normally, the voltages in the three standard bipolar limb leads, as measured from the peak of the R wave to the bottom of the S wave, vary between 0.5 and 2.0 mv, with lead III usually recording the lowest voltage and lead II the highest. However, these relationships are not invariably true even in the normal heart. In general, when the sum of the voltages of all the QRS complexes of the three standard leads is greater than 4 mv, one considers that the patient has a high-voltage electrocardiogram.

The cause of high voltage QRS complexes is most often increased muscular mass of the heart, which ordinarily results from *hypertrophy of the muscle* in response to excessive load on one part of the heart or the other. For instance, the right ventricle hypertrophies when it must pump blood through a stenotic pulmonary valve, and the left ventricle hypertrophies when a person has high blood pressure. The increased quantity of muscle allows generation of increased quantities of current around the heart. As a result, the electrical potentials recorded in the electrocardiographic leads are considerably greater than normal, as shown in Figures 16–13 and 16–14.

DECREASED VOLTAGE OF THE ELECTROCARDIOGRAM

There are three major causes of decreased voltage in the electrocardiogram. These are, first, abnormalities of the cardiac muscle itself that prevent generation of large quantities of current; second, abnormal conditions around the heart so that current cannot be conducted from the heart to the surface of the body with ease; and, third, rotation of the apex of the heart to point toward the anterior chest wall so that the electrical current of the heart flows mainly anteroposteriorly in the chest rather than in the frontal plane of the body, which decreases the voltages in the limb leads.

Decreased Voltage Caused by Cardiac Myopathies. One of the most usual causes of decreased voltage of the QRS complex is a series of *old myocardial infarctions* associated with *diminished muscle mass*. This also causes the depolarization wave to move through the ventricles slowly and prevents major portions of the heart from becoming massively depolarized all at once.

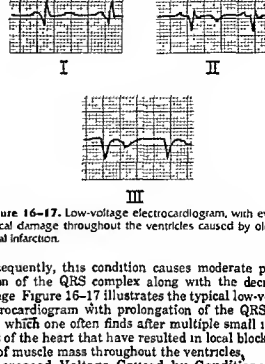


Figure 16–17. Low-voltage electrocardiogram, with evidence of local damage throughout the ventricles caused by old myocardial infarction.

Consequently, this condition causes moderate prolongation of the QRS complex along with the decreased voltage. Figure 16–17 illustrates the typical low-voltage electrocardiogram with prolongation of the QRS complex, which one often finds after multiple small infarctions of the heart that have resulted in local blocks and loss of muscle mass throughout the ventricles.

Decreased Voltage Caused by Conditions Surrounding the Heart. One of the most important causes of decreased voltage in the electrocardiographic leads is *fluid in the pericardium*. Because extracellular fluid conducts electrical currents with great ease, a large portion of current flowing out of the heart is conducted from one part of the heart to another through the pericardial effusion. Thus this effusion effectively "short-circuits" the current that is generated by the heart so that the proportion of this current reaching the surface of the body is greatly diminished. *Pleural effusion* to a lesser extent can also "short" the current around the heart so that the voltages at the surface of the body and in the electrocardiograms are decreased.

Pulmonary emphysema can decrease the electrocardiographic potentials but by a different method from that of pericardial effusion. In pulmonary emphysema, conduction of electrical current through the lungs is considerably depressed because of the excessive quantity of air in the lungs. Also, the chest cavity enlarges, and the lungs tend to envelop the heart to a greater extent than normally. Therefore, the lungs act as an insulator to prevent spread of current from the heart to the surface of the body, and this in general results in decreased electrocardiographic potentials in the various leads.

PROLONGED AND BIZARRE PATTERNS OF THE QRS COMPLEX

PROLONGED QRS COMPLEX AS A RESULT OF CARDIAC HYPERTROPHY OR DILATATION

The QRS complex lasts as long as the process of depolarization continues to spread through the ventricles—that is, as long as part of the ventricles is depolarized and part is still polarized. Therefore, the cause of a prolonged QRS complex is always *prolonged con-*

duction of the impulse through the ventricles. Such prolongation often occurs when one or both ventricles are hypertrophied or dilated, owing to the longer pathway that the impulse must then travel. The normal QRS complex lasts 0.06 to 0.09 second, whereas in hypertrophy or dilatation of the left or right ventricle the QRS complex may be prolonged to 0.10 to 0.12 second.

PROLONGED QRS COMPLEX RESULTING FROM PURKINJE SYSTEM BLOCKS

Block of the Purkinje fibers requires that the cardiac impulse be conducted by the ventricular muscle instead of through the specialized conduction system, thereby *decreasing the velocity of impulse conduction to approximately one-third to one-fourth normal.* Therefore, if complete block of one of the bundle branches occurs, the duration of the QRS complex is usually increased to 0.14 second or greater.

In general, a QRS complex is considered to be abnormally long when it lasts more than 0.09 second, and when it lasts more than 0.12 second the prolongation is almost certain to be caused by pathological block of the conduction system somewhere in the ventricles, as illustrated by the electrocardiograms for bundle branch block in Figures 16–15 and 16–16.

CONDITIONS CAUSING BIZARRE QRS COMPLEXES

Bizarre patterns of the QRS complex are most frequently caused by two conditions: first, destruction of cardiac muscle in various areas throughout the ventricular system with replacement of this muscle by scar tissue, and, second, local blocks in the conduction of impulses by the Purkinje system.

Sometimes local blocks occur in both the right and left ventricles. If the blocks are such that the impulse reaches the blocked area in the right ventricle much later than it reaches the blocked area in the left ventricle, but the total quantity of blocked muscle in the right ventricle is greater than the total quantity in the left ventricle, then a situation might occur in which the axis of the heart first shifts to the left and then, when the left ventricle has become totally depolarized but a portion of the right ventricle still remains polarized, rapidly shifts diametrically oppositely toward the right ventricle. This effect causes double or even triple peaks in some of the electrocardiographic leads, such as those illustrated in Figure 16–15.

CURRENT OF INJURY

Many different cardiac abnormalities, especially those that damage the heart muscle itself, often cause part of the heart to remain partially or totally *depolarized all the time.* When this occurs, current flows between the pathologically depolarized and the normally polarized [illegible] This is called a *current of injury.* Note especially *the injured part of the heart is negative because it is part that is depolarized and emits negative charges while the remainder of the heart is positive.*

Some of the abnormalities that can cause current of injury are: (1) *mechanical trauma,* which makes the membranes remain so permeable that full repolarization cannot take place; (2) *infectious processes* that damage the muscle membranes; and (3) *ischemia of local areas of muscle caused by coronary occlusion,* which is by far the most common cause of current of injury in the heart; during ischemia enough energy simply is not available to the heart muscle to maintain normal function.

EFFECT OF CURRENT OF INJURY ON THE QRS COMPLEX

In Figure 16–18 a shaded area in the base of the left ventricle is newly infarcted. Therefore, during the T-P interval—that is, when the normal ventricular muscle is polarized—current flows from the base of the left ventricle toward the rest of the ventricles. The vector of this "current of injury," as illustrated in the first and the last hearts, is in a direction of approximately 125 degrees, with the base of the vector, the *negative end,* toward the injured muscle. As illustrated in the lower portions of the figure, even before the QRS complex begins *this current flow causes an initial record in lead I below the zero potential line* because the projected vector of the current of injury in lead I points toward the negative end of the lead I axis. In lead II the record is above the line because the projected vector points toward the positive terminal of lead II. In lead III the projected vector of the current flow is also in the same *direction as the polarity of lead III so that the record is* positive. Furthermore, because the vector of the current of injury lies almost exactly along the axis of lead III, the potential of the current of injury in lead III is much greater than in either of the other two records.

As the heart then proceeds through its normal process of depolarization, the septum first becomes depolarized, and the depolarization spreads down to the apex and back toward the bases of the ventricles. The last portion of the ventricles to become totally depolarized is the base of the right ventricle, because the base of the left ventricle is already totally and permanently depolarized. By vectorial analysis, as illustrated in the figure, the electrocardiogram generated by the depolarization wave traveling through the ventricles may be constructed graphically, as shown in Figure 16–18.

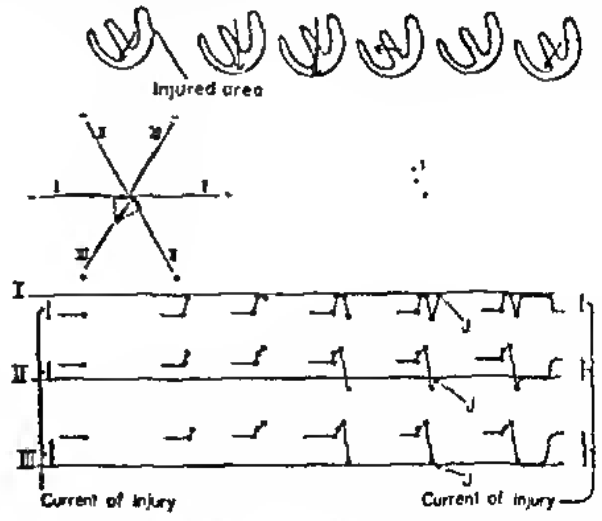


Figure 16–18. Effect of a current of injury on the electrocardiogram.

When the heart becomes totally depolarized at the end of the depolarization process, as noted by the next to last stage in Figure 16–18, all of the ventricular muscle is in a negative state. Therefore, at this instant in the electrocardiogram, absolutely no current flows around the musculature of the ventricles, because now the injured heart muscle and the contracting muscle are both completely depolarized

As repolarization then takes place, all the heart finally repolarizes except the area of permanent depolarization in the injured base of the left ventricle. Thus, the repolarization causes a return of the current of injury in each lead, as noted at the far right in Figure 16–18.

THE J POINT—THE ZERO REFERENCE POTENTIAL OF THE ELECTROCARDIOGRAM

One would think that the electrocardiograph machines for recording electrocardiograms could determine when no current is flowing around the heart However, many stray currents exist in the body, such as currents resulting from "skin potentials" and from differences in ionic concentrations in different parts of the body. Therefore, when two electrodes are connected between the arms or between an arm and a leg, these stray currents make it impossible for one to predetermine the exact zero reference level in the electrocardiogram. For these reasons, the following procedure must be used to determine the zero potential level. First, one notes *the exact point at which the wave of depolarization just completes its passage through the heart;* which occurs at the very end of the QRS complex. At exactly this point, all parts of the ventricles are depolarized so that no current is flowing around the heart. Even the current of injury disappears at this point. Therefore, the potential of the electrocardiogram at this instant is at exactly zero voltage. This point is known as the "J point" in the electrocardiogram, as illustrated in Figures 16–18 and 16–19.

For analysis of the electrical axis of the current of injury, a horizontal line is drawn through the electrocardiogram at the level of the J point, and this horizontal line is the zero potential line in the electrocardiogram from which all potentials caused by currents of injury must be measured.

Use of the J Point in Plotting the Axis of a Current of Injury. Figure 16–19 illustrates electrocardiograms recorded from leads I and III, both of which show currents of injury, In other words, the J point and the S-T segment of each of these two electrocardiograms are not on the same line as the T-P segment. A horizontal line has been drawn through the J point to represent the zero potential level in each of the two recordings. The potential of the current of injury in each lead is the difference between the level of the T-P segment of the electrocardiogram (which is recorded between heart beats when there is a current of injury) and the zero potential line, as illustrated by the arrows In lead I the recorded potential caused by the current of injury is above the zero potential line and is, therefore, positive. On the other hand, in lead III the T-P segment is below the zero potential line; therefore, the potential of the current of injury in lead III is negative.

At the bottom in Figure 16–19, the potentials of the current of injury in leads I and III are plotted on the coordinates of these leads, and the resultant vector of the current of injury for the whole ventricular mass is

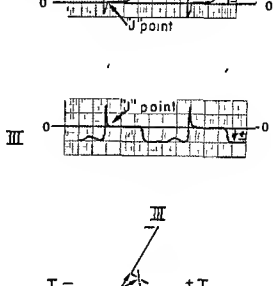


Figure 16–19. The "J point as the zero reference voltage of the electrocardiogram Also, method for plotting the axis of a current of injury is shown (*lower*)

determined by the method already described In this instance, the vector of the current of injury extends from the right side of the ventricles toward the left and slightly upward, with an axis of approximately −30 degrees

If one places the vector of the current of injury directly over the ventricles, *the negative end of the vector points toward the permanently depolarized, "injured" area of the ventricles.* In the instance illustrated in Figure 16–19, the injured area would be in the lateral wall of the right ventricle.

The Phenomenon of S-T Segment Shift. The portion *of the electrocardiogram that occurs between the* end of the QRS complex and the beginning of the T wave is called the *S-T segment.* The J point lies at the very beginning of this segment. Therefore, every time a current of injury occurs in one of the electrocardiographic leads, one also finds that the S-T segment and the T-P segments of the electrocardiogram are not at the same potential levels in the record. Actually, it is the T-P segment and not the S-T segment that is shifted away from the zero axis. However, most persons are conditioned to consider the T-P segment of the electrocardiogram as the reference potential level rather than the J point. Therefore, when a current of injury is evident in an electrocardiogram, it appears that the S-T segment is shifted from its normal level in the electrocardiogram, and this is called an *S-T segment shift.* Obviously, when one sees an S-T segment shift in an electrocardiogram, he or she knows immediately that the electrocardiogram shows the characteristics of a current of injury. In fact, most electrocardiographers do not speak of current of injury at all but simply speak of S-T segment shift, which means the same thing.

CORONARY ISCHEMIA AS A CAUSE OF CURRENT OF INJURY

Insufficient blood flow to the cardiac muscle depresses metabolism of the muscle for three different reasons: oxygen lack, excess carbon dioxide, and lack of sufficient nutrients. Consequently, repolarization of the membranes cannot occur in areas of severe myocardial ischemia. Often, the heart muscle does not die because the blood flow is sufficient to maintain life of the muscle even though it is not sufficient to cause repolarization of the membranes. As long as this state exists, a current of injury continues to flow during diastole.

Extreme ischemia of the cardiac muscle occurs following coronary occlusion, and a strong current of injury flows from the infarcted area of the ventricles during the period between heartbeats, as is illustrated in Figures 16–20 and 16–21. Therefore, one of the most important diagnostic features of electrocardiograms recorded following acute coronary thrombosis is the current of injury.

Acute Anterior Wall Infarction. Figure 16–20 illustrates the electrocardiogram in the three standard bipolar limb leads and in one chest lead recorded from a patient with acute anterior wall cardiac infarction. The most important diagnostic feature of this electrocardiogram is the intense current of injury in the chest lead. If one draws a zero potential line through the J point of this electrocardiogram, a strong *negative* current of injury during diastole is found, which means that the chest electrode over the front of the heart is in an area of strongly negative potential. In other words, the negative end of the current of injury vector is against the chest wall. This means that the current of injury is emanating from the anterior wall of the ventricles, which is the main reason for diagnosing this condition as anterior wall infarction.

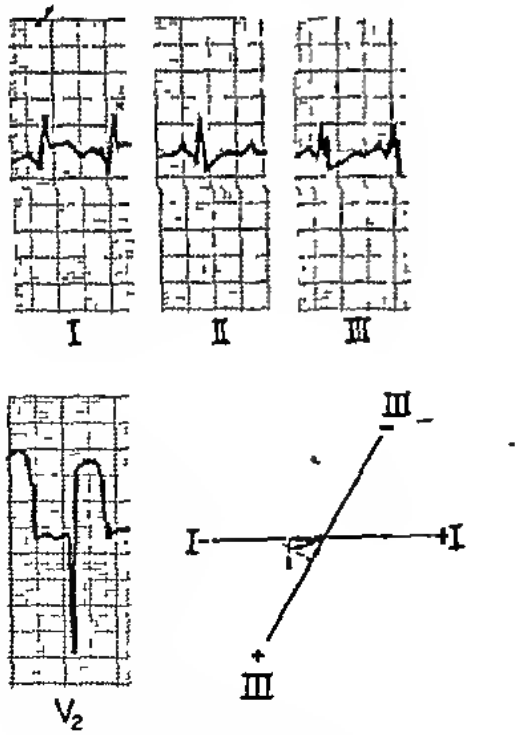


Figure 16–20. Current of injury in acute anterior wall infarction. The intense current of injury in lead V_2.

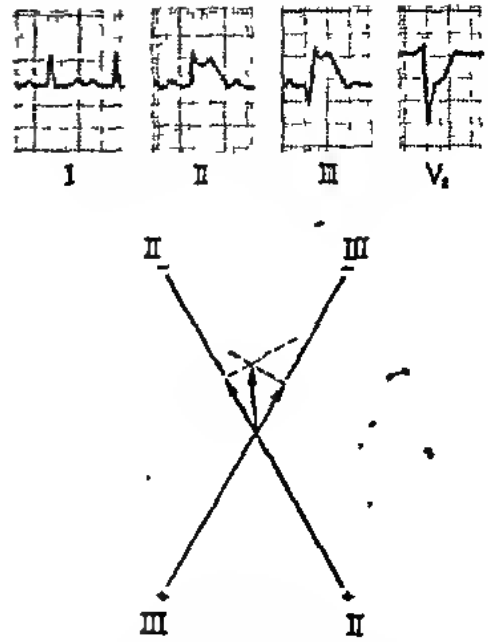


Figure 16–21. Current of injury in acute posterior wall, apical infarction.

Analyzing the currents of injury in leads I and III, one finds a negative potential caused by the current of injury in lead I and a positive potential for the current of injury in lead III. This means that the resultant vector of the current of injury in the heart is approximately +150 degrees, with the negative end of the vector pointing toward the left ventricle and the positive end pointing toward the right ventricle. Thus, in this particular electrocardiogram the current of injury appears to be coming mainly from the left ventricle and from the anterior wall of the heart. Therefore, one would suspect that this anterior wall infarction is probably caused by thrombosis of the anterior descending limb of the left coronary artery.

Posterior Wall Infarction. Figure 16–21 illustrates the three standard bipolar limb leads and one chest lead from a patient with posterior wall infarction. The major diagnostic feature of this electrocardiogram is also in the chest lead. If a zero potential reference line is drawn through the J point of this lead, it is readily apparent that during the T-P interval the potential of the current of injury is positive. This means that the positive end of the vector is at the chest wall and the negative end (injured end) is away from the chest wall. In other words, the current of injury is coming from the side of the heart opposite that portion adjacent to the chest wall, which is the reason why this type of electrocardiogram is the basis for diagnosing posterior wall infarction.

If one analyzes the currents of injury in leads II and III of Figure 16–21, it is readily apparent that the current of injury is negative in both leads. By vectorial analysis as shown in the figure, one finds that the vector of the current of injury is approximately −95 degrees, with the negative end of the vector pointing downward and the positive end upward. Thus, because the infarct, as indicated by the chest lead, is on the posterior wall of the heart and, as indicated by the currents of injury in leads II and III, is in the apical portion of the heart, one would suspect that this infarct is close to the apex on the posterior wall of the left ventricle.

Infarction in Other Parts of the Heart. By the same procedures as those illustrated in the preceding two discussions of anterior and posterior wall infarctions, it is possible to determine the locus of an infarcted area emitting a current of injury regardless of which part of the heart is involved. In making such vectorial analyses, it must always be remembered that *the positive end of the vector points toward the normal cardiac muscle and the negative end points toward the abnormal portion of the heart that is emitting the current of injury.*

Recovery From Coronary Thrombosis. Figure 16–22 illustrates a V_3 chest lead from a patient with posterior infarction, showing the change in the electrocardiogram of this lead from the day of the attack to one week later, then three weeks later, and finally one year later. From this electrocardiogram it can be seen that the current of injury is strong immediately after the acute attack (T-P segment displaced positively from the J point and the S-T segment), but after approximately one week the current of injury has diminished considerably and after three weeks it is completely gone. After that, the electrocardiogram does not change greatly during the following year. This is the usual recovery pattern following cardiac infarction of moderate degree when the collateral coronary blood flow is sufficient to reestablish appropriate nutrition to most of the infarcted area.

On the other hand, when all the coronary vessels throughout the heart are fairly well sclerosed, it may not be possible for the adjacent coronary vessels to supply sufficient blood to the infarcted area for recovery. Therefore, in some patients with coronary infarction, the infarcted area never redevelops an adequate coronary blood supply, some of the heart muscle dies, and relative coronary insufficiency persists in this area of the heart indefinitely. If the muscle does not die and become replaced by scar tissue, it continually emits a current of injury as long as the relative ischemia exists, particularly during bouts of exercise when the heart is overloaded.

Old Recovered Myocardial Infarction. Figure 16–23 illustrates leads I and III following anterior infarction and posterior infarction as these leads appear approximately a year after the acute episode. These are what might be called the "ideal" configurations of the QRS complex in these types of recovered myocardial infarction. It is illustrated in these figures that usually a Q wave develops at the beginning of the QRS complex in lead I in anterior infarction because of loss of muscle mass in the anterior wall of the left ventricle; whereas

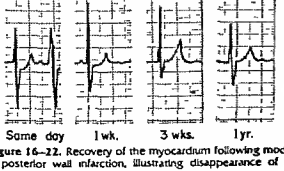


Figure 16–22. Recovery of the myocardium following moderate posterior wall infarction, illustrating disappearance of the current of injury (lead V_3).

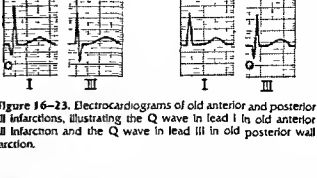


Figure 16–23. Electrocardiograms of old anterior and posterior wall infarctions, illustrating the Q wave in lead I in old anterior wall infarction and the Q wave in lead III in old posterior wall infarction.

in posterior infarction a Q wave develops at the beginning of the QRS complex in lead III because of loss of muscle in the posterior apical part of the ventricle.

These configurations are certainly not those found in all cases of old anterior and posterior cardiac infarction. Local loss of muscle and local areas of conduction block can cause the following abnormalities of the QRS complex: bizarre patterns (the prominent Q waves, for instance), decreased voltage, and prolongation.

Current of Injury in Angina Pectoris. "Angina pectoris" means simply pain in the pectoral regions of the upper chest, this pain usually radiating into the neck and down the left arm. The pain is caused by *relative* ischemia of the heart. No pain is usually felt as long as the person is perfectly quiet, but, just as soon as the patient overworks the heart, the pain appears.

A current of injury often occurs during an attack of severe angina pectoris, for the relative coronary insufficiency then occasionally becomes great enough to prevent adequate repolarization of the membranes in some areas of the heart during diastole.

ABNORMALITIES IN THE T WAVE

Earlier in the chapter it was pointed out that the T wave is normally positive in all the standard bipolar limb leads and that this is caused by repolarization of the apex of the heart ahead of the endocardial surfaces of the ventricles. This direction in which repolarization spreads over the heart is backward to the direction in which depolarization takes place. (If the basic principles of the upright T wave in the standard leads have not been understood by now, the student should become familiar with the earlier, more detailed discussion of this before proceeding to the following few sections.)

The T wave becomes abnormal when the normal sequence of repolarization does not occur. Several factors can change this sequence of repolarization, as follows.

EFFECT OF SLOW CONDUCTION OF THE DEPOLARIZATION WAVE ON THE T WAVE

Referring back to Figure 16–15, note that the QRS complex is considerably prolonged. The reason for this prolongation is delayed conduction in the left ventricle as a result of left bundle branch block. The left ventricle becomes depolarized approximately 0.08 second after depolarization of the right ventricle, which gives a strong mean QRS vector to the left. The refractory periods of the right and left ventricular muscle masses are not greatly different from each

right ventricle begins to repolarize long before the left ventricle; this causes positivity in the right ventricle and negativity in the left ventricle. In other words, the mean axis of the T wave is *to the right*, which is opposite the mean electrical axis of the QRS complex in this same electrocardiogram. Thus, when conduction of the impulse through the ventricles is greatly delayed, the T wave is almost always of opposite polarity to that of the QRS complex.

In Figure 16-16 and in several figures in the following chapter, conduction also does not occur through the Purkinje system. As a result, the rate of conduction is greatly slowed, and in each instance the T wave is of opposite polarity to that of the QRS complex, whether the condition causing this delayed conduction happens to be left bundle branch block, right bundle branch block, premature ventricular contraction, or otherwise.

PROLONGED DEPOLARIZATION IN PORTIONS OF THE VENTRICULAR MUSCLE AS A CAUSE OF ABNORMALITIES IN THE T WAVE

If the apex of the ventricles should have an abnormally long period of depolarization, that is, a prolonged action potential, then repolarization of the ventricles would not begin at the apex as it normally does. Instead, the base of the ventricles would repolarize ahead of the apex, and the vector of repolarization would point from the apex toward the base of the heart. This is opposite to the usual vector of repolarization, and, consequently, the T wave in all three standard leads would be negative rather than the usual positive. Thus, the simple fact that the apical muscle of the heart had a prolonged period of depolarization would be sufficient to cause marked changes in the T wave, even to the extent of changing the entire polarity, as is illustrated in Figure 16-24.

Ischemia is by far the most common cause of increased duration of depolarization of cardiac muscle, and, when the ischemia occurs in only one area of the heart, the depolarization period of this area increases out of proportion to that in other portions. As a result, definite changes in the T wave may take place. The ischemia may result from chronic, progressive coronary occlusion, acute coronary occlusion, or relative coronary insufficiency occurring during exercise.

One means for detecting mild coronary insufficiency is to have the patient exercise and then to record the electrocardiogram immediately thereafter, noting whether or not changes occur in the T waves. The changes in the T waves need not be specific, for any change in the T wave in any lead—inversion, for instance, or a biphasic wave—is often evidence enough that some portion of the ventricular muscle has increased its period of depolarization out of proportion to the rest of the heart, and this is probably caused by relative coronary insufficiency.

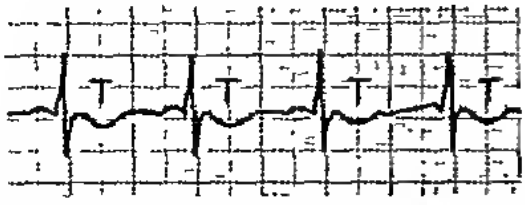


Figure 16-24. Inverted T wave resulting from mild ischemia of the apex of the ventricles.

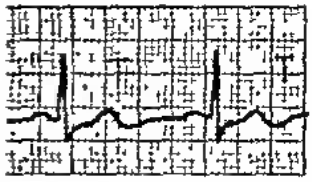


Figure 16-25. Biphasic T wave caused by digitalis toxicity.

All the other conditions that can cause currents of injury, including pericarditis, myocarditis, and mechanical trauma of the heart, can also cause changes in the T wave. A current of injury occurs when the period of depolarization of some muscle is so long that the muscle fails to repolarize completely before the next cardiac cycle begins. Therefore, a current of injury is actually an exacerbated form of abnormal T wave, for both of these result from an increased depolarization period of one or more portions of cardiac muscle, and the difference is only a matter of degree.

Effect of Digitalis on the T Wave. As discussed in Chapter 26, digitalis is a drug that can be used during relative coronary insufficiency to increase the strength of cardiac muscular contraction. However, digitalis also increases the period of depolarization of cardiac muscle. It usually increases this period by approximately the same proportion in all or most of the ventricular muscle, but when overdosages of digitalis are given, the depolarization period of one part of the heart may be increased out of proportion to that of other parts. As a result, nonspecific changes, such as T wave inversion or biphasic T waves, may occur in one or more of the electrocardiographic leads. A biphasic T wave caused by excessive administration of digitalis is illustrated in Figure 16-25. There is a slight amount of current of injury, too. This probably results from continuous depolarization of part of the ventricular muscle.

Changes in the T wave during digitalis administration are the earliest signs of digitalis toxicity. If still more digitalis is given to the patient, strong currents of injury may develop, and, as noted in the following chapter, digitalis can also block conduction of the cardiac impulse to various portions of the heart, so that various arrhythmias can result. It is desirable clinically to keep the effects of digitalis from going beyond the stage of mild T wave abnormalities. Therefore, the electrocardiograph is used routinely in following digitalized patients.

REFERENCES

See bibliography at end of Chapter 17

17

Electrocardiographic Interpretation of Cardiac Arrhythmias

The rhythmicity of the heart and some abnormalities of rhythmicity were discussed in Chapter 14. The major purpose of the present chapter is to discuss still other abnormalities of rhythm and to describe the electrocardiograms recorded in the conditions known clinically as "cardiac arrhythmias."

ABNORMAL SINUS RHYTHMS

TACHYCARDIA

The term "tachycardia" means fast heart rate, usually defined as faster than 100 beats per minute. An electrocardiogram recorded from a patient with tachycardia is illustrated in Figure 17–1. This electrocardiogram is normal except that the rate of heartbeat, as determined from the time intervals between QRS complexes, is approximately 150 per minute instead of the normal 72 per minute.

The three general causes of tachycardia are *increased body temperature, stimulation of the heart by the sympathetic nerves,* and *toxic conditions of the heart.*

The rate of the heart increases approximately 10 beats per minute for each degree Fahrenheit (18 beats per degree Celsius) increase in body temperature up to a body temperature of about 105°; beyond this the heart rate may actually decrease, owing to progressive weakening of the heart muscle as a result of the fever. Fever causes tachycardia because increased temperature increases the rate of metabolism of the S-A node, which in turn directly increases its excitability and rate of rhythm.

The various factors that can cause the sympathetic nervous system to excite the heart will be discussed in Chapter 22 in relation to the reflexes that control the circulatory system. For instance, when a patient loses blood and passes into a state of shock or semishock, reflex stimulation of the heart increases its rate to as high as 150 to 180 beats per minute. Also, simple weakening of the myocardium usually increases the heart rate because the weakened heart does not pump blood into the arterial tree to a normal extent, and this elicits reflexes to increase the rate of the heart.

BRADYCARDIA

The term "bradycardia" means a slowed heart rate, usually defined as less than 60 beats per minute. Bradycardia is illustrated by the electrocardiogram in Figure 17–2.

Bradycardia in Athletes. The athlete's heart is considerably stronger than that of a normal person, a fact that allows the athlete's heart to pump a greater stroke volume output per beat. The excessive quantities of blood pumped into the arterial tree with each beat presumably initiate circulatory reflexes to cause the bradycardia.

Vagal Stimulation as a Cause of Bradycardia. Obviously, any circulatory reflex that stimulates the vagus nerve can cause the heart rate to decrease considerably. Perhaps the most striking example of this occurs in patients with the *carotid sinus syndrome.* In these patients an arteriosclerotic process in the carotid sinus region of the carotid artery causes excessive sensitivity of the pressure receptors (baroreceptors) located in the arterial wall; as a result, mild pressure on the neck elicits a *strong baroreceptor reflex, causing intense vagal stimulation* of the heart and extreme bradycardia. Indeed, sometimes this reflex is so powerful that it actually stops the heart.

SINUS ARRHYTHMIA

Figure 17–3 illustrates a *cardiotachometer* recording of the heart rate during normal and deep respiration. A

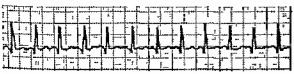


Figure 17–1. Sinus tachycardia (lead I)

Figure 17–2. Sinus bradycardia (lead III)

Figure 17–3. Sinus arrhythmia as detected by a cardiotachometer. To the left is the recording taken when the subject was breathing normally, to the right, when breathing deeply.

cardiotachometer is an instrument that records by the height of successive spikes the duration of the interval between each two QRS complexes in the electrocardiogram. Note from this record that the heart rate increases and decreases approximately 5 per cent during the various phases of the quiet respiratory cycle. However, during deep respiration, as shown to the right in Figure 17–3, the heart rate even normally increases and decreases by as much as 30 per cent.

Sinus arrhythmia is the result of several circulatory reflexes that will be described in detail in Chapter 22. First, when the blood pressure rises and falls during each cycle of respiration, the *baroreceptors are alternately stimulated and depressed*, causing reflex slowing and speeding up of the heart. Second, during each respiratory cycle, the negative intrapleural pressure increases and decreases, which increases and decreases the effective pressure in the veins in the chest and in the right atrium. This elicits a *waxing and waning Bainbridge reflex* that also increases and decreases the cardiac rate. Third, when the respiratory center of the medulla is excited during each respiratory cycle, some of the impulses *"spill over" from the respiratory center into the vasomotor center*, causing alternate increase and decrease in the number of impulses transmitted to the heart through the sympathetic and vagus nerves.

ABNORMAL RHYTHMS RESULTING FROM IMPULSE CONDUCTION BLOCK

SINOATRIAL BLOCK

In rare instances the impulse from the S-A node is blocked before it enters the atrial muscle. This phenomenon is illustrated in Figure 17–4, which shows the sudden cessation of P waves with resultant standstill of the atrium. However, the ventricle picks up a new rhythm, the impulse usually originating in the A-V node, so that the ventricular QRS-T complex is not altered.

ATRIOVENTRICULAR BLOCK

The only means by which impulses can ordinarily pass from the atria into the ventricles is through the *A-V bundle*, which is also known as the *bundle of His*. The different conditions that can either decrease the rate of conduction of the impulse through this bundle or can totally block the impulse are:

1. *Ischemia of the A-V nodal fibers* often delays or blocks conduction from the atria to the ventricles. Coronary insufficiency can cause ischemia of the A-V node in the same manner that it can cause ischemia of the myocardium.

2. *Compression of the A-V bundle* by scar tissue or by calcified portions of the heart can depress or block conduction from the atria to the ventricles.

3. *Inflammation of the A-V node or A-V bundle* can depress conductivity between the atria and the ventricles. Inflammation results frequently from different types of myocarditis such as occur in diphtheria and rheumatic fever.

4. *Extreme stimulation of the heart by the vagus nerves* in rare instances blocks impulse conduction through the A-V node. Such vagal excitation occasionally results from strong stimulation of the baroreceptors in persons with the *carotid sinus syndrome*, which was discussed earlier in relation to bradycardia.

Incomplete Heart Block. *Prolonged P-R (or P-Q) Interval—"First Degree Block."* The normal lapse of time between the *beginning* of the P wave and the *beginning* of the QRS complex is approximately 0.16 second when the heart is beating at a normal rate. This P-R interval usually decreases in length with faster heartbeat and increases with slower heartbeat. In general when the P-R interval increases above a value of approximately 0.20 second in the heart beating at normal rate, the P-R interval is said to be prolonged, and the patient is said to have *first degree incomplete heart block*. Figure 17–5 illustrates an electrocardiogram with a prolonged P-R interval, the interval in this instance being approximately 0.30 second.

The P-R interval rarely increases above 0.35 to 0.45 second, for by the time the rate of conduction through the A-V node and bundle is depressed to such an extent, conduction stops entirely. Thus, when a patient's P-R interval is approaching these limits, additional slight increase in the severity of the condition will block impulse conduction rather than simply delay conduction further.

One of the means for determining the severity of some heart diseases—rheumatic fever, for instance—is to measure the P-R interval.

Second Degree Block. When conduction through the A-V junction is slowed until the P-R interval is 0.25 to 0.45 second, the action potentials traveling through the A-V node sometimes are strong enough to pass on through the A-V node and at other times are not strong enough. Often the impulse passes into the ventricles following one atrial contraction and fails to pass following the next one or two atrial contractions, thus alternating between conduction and nonconduction. In this instance, the atria beat at a considerably faster rate than the ventricles, and it is said that there are "dropped beats" of the ventricles. This condition is called *second degree incomplete heart block*.

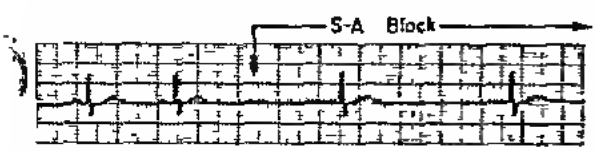


Figure 17–4. S-A nodal block with A-V nodal rhythm (lead III).

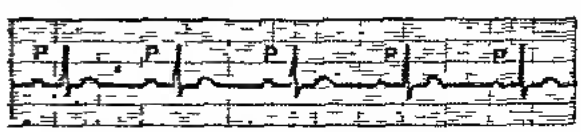


Figure 17–5. Prolonged P-R interval (lead II).

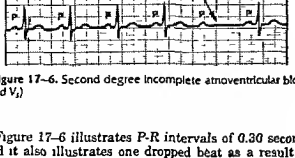


Figure 17–6. Second degree incomplete atrioventricular block (lead V_1)

Figure 17–6 illustrates P-R intervals of 0.30 second, and it also illustrates one dropped beat as a result of failure of conduction from the atria to the ventricles

At times every other beat of the ventricles is dropped so that a "2:1 rhythm" develops in the heart, with the atria beating twice for every single beat of the ventricles Sometimes other rhythms, such as 3 2 or 3.1, also develop.

Complete Atrioventricular Block (Third Degree Block). When the condition causing poor conduction in the A-V node or bundle becomes extremely severe, complete block of the impulse from the atria into the ventricles occurs. In this instance tha P waves become completely dissociated from the QRS-T complexes, as illustrated in Figure 17–7. Note that the rate of rhythm of the atria in this electrocardiogram is approximately 100 beats per minute, while the rate of ventricular beat is less than 40 per minute Furthermore, there is no relationship whatsoever between the rhythm of the atria and that of the ventricles, for the ventricles have "escaped" from control by the atria, and they are beating at their own natural rate.

Stokes-Adams Syndrome—Ventricular Escape In some patients with atrioventricular block, the total block comes and goes—that is, impulses are conducted from the atria into the ventricles for a period of tima, and then suddenly no impulses at all are transmitted. Tha duration of total block may be a few seconds, a few minutes, a few hours, or it may be weeks or even longer before conduction returns. In particular, this condition occurs in hearts with borderline ischemia.

Immediately after A-V conduction is first blocked, the ventricles stop contracting entirely for 5 to 30 seconds because of the phenomenon called *overdrive suppression* that was discussed in Chapter 14. Then some part of the Purkinje system beyond the block, usually in the distal part of the A-V node beyond the blocked point in the node or in the A-V bundle, begins discharging rhythmically at a rate of 15 to 40 times per minute and acting as the pacemaker of the ventricles. This is called *ventricular escape.* Because the brain cannot remain active for more than 3 to 5 seconds without blood supply, patients usually faint a few seconds after complete block occurs because the heart then does not pump any blood for 5 to 30 seconds until the ventricles "escape." However, even the slowly beating ventricles usually pump enough blood to allow rapid recovery from the faint and then to sustain the person. These periodic fainting spells are known as the Stokes-Adams syndrome.

Occasionally the interval of ventricular standstill at the onset of complete block is so long that it either becomes detrimental to the health of the patient or causes death. Consequently, many of these patients would be better off with total heart block than constantly shifting back and forth between total heart block and atrial control of the ventricles. Such patients are frequently treated with drugs that depress conduction in the A-V node so that the total heart block persists all the time. Commonly, digitalis is the drug used for this purpose, but quinidine or barium ion can also convert a Stokes-Adams syndrome into total heart block.

However, most of these patients are now provided with an *artificial pacemaker,* which is a small battery-operated electrical stimulator planted beneath the skin, the electrodes from which are connected to the right ventricle. This pacemaker provides continued rhythmic impulses that take over control of the ventricles. The batteries are replaced about once every five years.

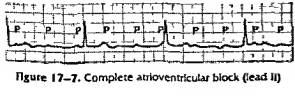


Figure 17–7. Complete atrioventricular block (lead II)

INCOMPLETE INTRAVENTRICULAR BLOCK—ELECTRICAL ALTERNANS

Most of the same factors that can cause A-V block can also block impulse conduction in peripheral portions of tha ventricular Purkinje system. At times, *incomplete block occurs so that the impulse is sometimes transmitted* and sometimes not transmitted, causing block of the impulse during some heart cycles and not during others. Tha QRS complex may be considerably abnormal during those cycles in which the impulse is blocked. Figure 17–8 illustrates the condition known as *electrical alternans,* which results from partial intraventricular block every other heartbeat. This electrocardiogram also illustrates tachycardia, which is probably the reason the block has occurred, for when the rate of the heart is very rapid it may be impossible for portions of the Purkinje system to recover from the refractory period quickly enough to respond during each succeeding heartbeat. Also, many conditions that depress the heart, such as ischemia, myocarditis, and digitalis toxicity, can cause incomplete intraventricular block with resultant electrical alternans.

PREMATURE CONTRACTIONS

A premature contraction is a contraction of the heart prior to the time that normal contraction would have been expected. This condition is also frequently called *extrasystole* or *premature beat.*

Causes of Premature Contractions. Most premature contractions result from *ectopic foci* in the heart,

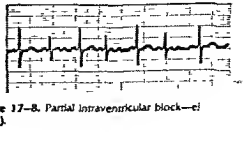


Figure 17–8. Partial intraventricular block—el (lead III).

which emit abnormal impulses at odd times during the cardiac rhythm. The possible causes of ectopic foci are (1) local areas of ischemia, (2) small calcified plaques at different points in the heart, which press against the adjacent cardiac muscle so that some of the fibers are irritated, and (3) toxic irritation of the A-V node, Purkinje system, or myocardium caused by drugs, nicotine, or caffeine. Mechanical initiation of premature contractions is also frequent during cardiac catheterization, large numbers of premature contractions often occurring when the catheter enters the right ventricle and presses against the endocardium.

In persons with ischemic heart disease, the ectopic rhythm is, on occasion, believed to be caused by a reentrant signal as follows. The normal heart beat excites an area of ischemic tissue in which the cardiac impulse travels extremely slowly. Then, after the contraction of the normal heart muscle is over, the slowly traveling signal in the ischemic tissue escapes back into the normal cardiac muscle, in this way causing a second contraction at a later time in the heartbeat.

PREMATURE ATRIAL CONTRACTIONS

Figure 17–9 is an electrocardiogram showing a single premature atrial contraction. The P wave of this beat is relatively normal and the QRS complex is also normal, but the P-R interval and the interval between the preceding contraction and the premature contraction is shortened. Also, the interval between the premature contraction and the next succeeding contraction is slightly prolonged, which is called a *compensatory pause*. The reason for this is that the premature contraction originated in the atrium some distance from the S-A node, and the impulse from its origin had to travel through a considerable amount of atrial muscle before it discharged the S-A node. Consequently, the S-A node discharged very late in the premature cycle, and this made the succeeding heartbeat also late in appearing.

Premature atrial contractions occur frequently in healthy persons and, indeed, are often found in athletes or others whose hearts are certain to be in healthy condition. Yet mild toxic conditions resulting from such factors as excess smoking, lack of sleep, too much coffee, alcoholism, and use of various drugs can also initiate such contractions.

Pulse Deficit. When the heart contracts ahead of schedule, the ventricles will not have filled with blood normally, and the stroke volume output during that contraction is depressed or sometimes almost absent. Therefore, the pulse wave passing to the periphery following a premature contraction may be so weak that the pulse cannot be felt at all in the radial artery, whereas the succeeding pulse may be extra strong because of compensatory overfilling. Thus, a deficit in the number of pulses felt in the radial pulse occurs in relation to the number of contractions of the heart.

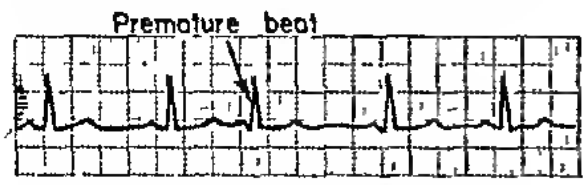


Figure 17–9. Atrial premature contraction (lead I).

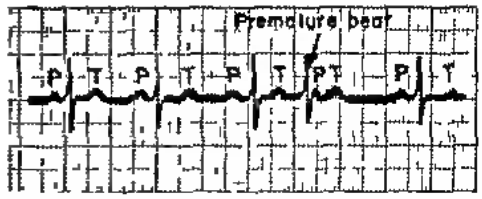


Figure 17–10. A-V nodal premature contraction (lead III).

Bigeminal Pulse. At times every other beat of the heart may be a premature contraction. This causes the patient to have a bigeminal pulse—that is, two pulses close together, then a longer diastolic interval, then two again, and so on.

A-V NODAL OR A-V BUNDLE PREMATURE CONTRACTIONS

Figure 17–10 illustrates a premature contraction originating either in the A-V node or in the A-V bundle. The P wave is missing from the record of the premature contraction. Instead, the P wave is superimposed onto the QRS-T complex of the premature contraction because the cardiac impulse travels backward into the atria at the same time that it travels forward into the ventricles; this P wave distorts the complex, but the P wave itself cannot be discerned as such.

In general, A-V nodal premature contractions have the same significance and causes as atrial premature contractions.

PREMATURE VENTRICULAR CONTRACTIONS (PVCs)

The electrocardiogram of Figure 17–11 illustrates a series of premature ventricular contractions alternating with normal contractions. Most premature ventricular

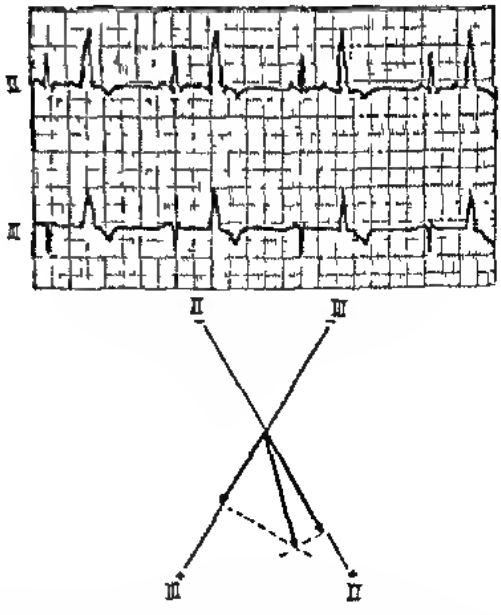


Figure 17–11. Premature ventricular contractions (PVCs) illustrated by the large abnormal QRS-T complexes (leads II and III). Axis of the premature contractions is plotted in accord with the principles of vectorial analysis explained in Chapter 16.

contractions of this type probably result from a re-entrant signal from an ischemic area of ventricular muscle, as described earlier. This causes several effects in the electrocardiogram:

First, the QRS complex is usually considerably prolonged. The reason is that the impulse is conducted mainly through the slowly conducting muscle of the ventricle rather than through the Purkinje system.

Second, the QRS complex has a very high voltage for the following reason: When the normal impulse passes through the heart, it passes through both ventricles approximately simultaneously; consequently, the depolarization waves of the two sides of the heart partially neutralize each other. However, when a ventricular premature contraction occurs, the impulse travels in only one direction so that there is no such neutralization effect, and one entire side of the heart is depolarized while the other entire side is still polarized; this causes intense electrical potentials

Third, following almost all premature ventricular contractions the T wave has a potential opposite to that of the QRS complex, because the *slow conduction of the impulse* through the cardiac muscle causes the area first depolarized also to repolarize first. As a result, the direction of current flow in the heart during repolarization is opposite to that during depolarization, and the potential of the T wave is reversed to that of the QRS complex. This is not true of the normal T wave, as was explained in Chapter 16.

Some premature ventricular contractions are relatively benign in their origin and result from simple factors such as cigarettes, coffee, lack of sleep, various mild toxic states, and even emotional irritability. On the other hand, a large share of premature ventricular contractions results from stray impulses or re-entrant signals originating around the borders of infarcted or ischemic areas of the heart. Therefore, the presence of such PVCs is not to be taken lightly. Statistics show that persons with significant numbers of such PVCs have a much higher than normal chance of developing spontaneous lethal ventricular fibrillation, presumably initiated by one of the PVCs itself. This is especially true when the PVCs occur during the vulnerable period for causing fibrillation, just at the end of the T wave when the ventricles are coming out of refractoriness, as explained in Chapter 14.

Vector Analysis of the Origin of an Ectopic Premature Ventricular Contraction. In Chapter 16 the principles of vectorial analysis were explained. Applying these principles, one can determine from the electrocardiogram in Figure 17–11 the point of origin of the premature ventricular contraction as follows: Note that the potentials of the premature contractions in leads II and III are both strongly positive. Plotting these potentials on the axes of leads II and III and solving by vectorial analysis for the resultant mean QRS vector in the heart, one finds that this vector of the premature contraction has its negative end (origin) at the base of the heart and its positive end toward the apex. Thus, the first portion of the heart to become depolarized during the premature contractions lies near the base of the heart, which therefore is the locus of the ectopic focus.

PAROXYSMAL TACHYCARDIA

Abnormalities in any portion of the heart, including the atria, the Purkinje system, and the ventricles, can sometimes cause rapid rhythmic discharge of impulses which spread in all directions throughout the heart. This is most frequently caused by re-entrant pathways that set up local circus movements. Because of the rapid rhythm in the irritable focus, this focus becomes the pacemaker of the heart.

The term "paroxysmal" means that the heart rate usually becomes very rapid in paroxysms, with the paroxysms beginning suddenly and lasting for a few seconds, a few minutes, a few hours, or sometimes much longer. Then the paroxysms usually end as suddenly as they had begun, the pacemaker of the heart then shifting back to the S-A node.

Paroxysmal tachycardia can often be stopped by eliciting a vagal reflex. A strange type of vagal reflex sometimes elicited for this purpose is one that occurs when painful pressure is applied to the eyes. Also, pressure on the carotid sinuses can sometimes elicit enough of a vagal reflex to stop the tachycardia. Various drugs may also be used. Two frequently used are quinidine and lidocaine, both of which depress the normal increase in sodium permeability of the cardiac muscle membrane during the generation of the action potential, thereby often blocking the rhythmic discharge of the focal region that is causing the paroxysmal attack.

ATRIAL PAROXYSMAL TACHYCARDIA

Figure 17–12 illustrates in the middle of the record a sudden increase in rate of heartbeat from approximately 95 to approximately 150 beats per minute. On close observation of the electrocardiogram it will be seen that an inverted P wave occurs before each of the QRS-T complexes during the paroxysm of rapid heartbeat, and this P wave is partially superimposed on the normal T wave of the preceding beat. This indicates that the origin of this paroxysmal tachycardia is in the atrium, but, because the P wave is abnormal, the origin is not near the S-A node.

A-V Nodal Paroxysmal Tachycardia. Paroxysmal tachycardia very often results from an aberrant rhythm that involves the A-V node. The usual cause is a local re-entrant circus movement. Two of the most common of these are the following. (1) Occasionally, some of the A-V nodal fibers are very slow to come out of refractoriness and others come out rapidly. Therefore, after ventricular contraction, occasionally a re-entrant cardiac impulse is transmitted backward through the rapidly recovering A-V nodal fibers. This then re-excites the proximal portions of the A-V node (as well as the atrium itself) and the signal returns once again to the ventricles, but through the slow A-V nodal fibers that have by now come out of the refractory state. Thus a new ventricular contraction occurs, followed by re-entry once again through the rapidly recovering A-V nodal fibers and continuation of the repetitive cycle.

(2) In some hearts a second bundle of cardiac fibers besides the A-V bundle connects the atrial and ventricular muscle masses. This extra muscle pathway is frequently called the *bundle of Kent*. In such instances, the

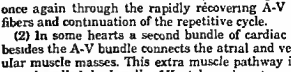


Figure 17–12. Atrial paroxysmal tachycardia—onset in middle of record (lead I).

signal transmitted through the A-V bundle into the ventricles can return to the atria through the bundle of Kent and cause a second contraction shortly after the first. Or it is also possible for the signal to go first through the bundle of Kent and then to return by way of the A-V bundle. In either event a local circus movement is initiated.

Regardless of the cause of the A-V nodal tachycardia, it usually causes almost normal QRS-T complexes.

Atrial or A-V nodal paroxysmal tachycardia, both of which are called *supraventricular tachycardias*, usually occurs in young, otherwise healthy persons. In general, supraventricular tachycardia frightens the individual tremendously and may cause weakness during the paroxysms but only rarely does permanent harm come from the attacks.

VENTRICULAR PAROXYSMAL TACHYCARDIA

Figure 17–13 illustrates a typical short paroxysm of ventricular tachycardia. The electrocardiogram of ventricular paroxysmal tachycardia has the appearance of a series of ventricular premature beats occurring one after another without any normal beats interspersed.

Ventricular paroxysmal tachycardia is usually a serious condition for two reasons. First, this type of tachycardia usually does not occur unless considerable ischemic damage is present in the ventricles. Second, ventricular tachycardia very frequently initiates ventricular fibrillation, which is almost invariably fatal. The reasons why ventricular tachycardia predisposes to ventricular fibrillation are: First, rapid stimulation decreases the refractory period of the heart and slows conduction, both of which predispose to circus movements. Second, one of the ventricular impulses frequently occurs at an instant when part of the ventricular muscle is still in a refractory state and part in a nonrefractory state—that is, during the vulnerable period for initiating fibrillation, this causes the impulse to take a devious, one-way route through the heart, avoiding the refractory areas and initiating a long circuitous pathway, as was discussed in Chapter 14. The long pathway allows the impulse to continue around and around the ventricles, resulting in ventricular fibrillation.

Digitalis intoxication sometimes causes ventricular tachycardia. On the other hand, quinidine or lidocaine, which increases the refractory period of cardiac muscle and also increases its threshold for excitation, may be used to block irritable foci causing ventricular tachycardia.

FLUTTER AND FIBRILLATION

As explained in Chapter 14, when a circus movement spreads asynchronously around a long pathway in the atria, *atrial flutter* occurs. On the other hand, when the circus movement becomes asynchronous in either the atrial muscle or the ventricular muscle, fibrillation occurs—*atrial fibrillation* in the first instance, *ventricular fibrillation* in the second instance.

Figure 17–13. Ventricular paroxysmal tachycardia (lead III).

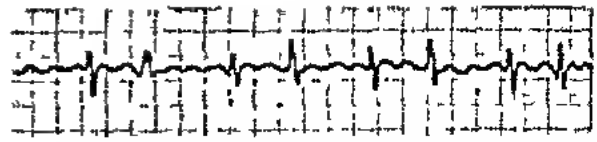


Figure 17–14. Atrial flutter—2 1 and 3 1 rhythm (lead I).

ATRIAL FLUTTER

Figure 17–14 illustrates lead II of the electrocardiogram in atrial flutter. The rate of atrial contraction (P waves) is approximately 300 times per minute, while the rate of ventricular contraction (QRS-T waves) is only 125 times per minute. From the record it will be seen that sometimes either a 2:1 or a 3:1 rhythm occurs. In other words, because of the long refractory period of the A-V node the atria contract two or three times for every one impulse that is conducted through the A-V bundle into the ventricles.

An important distinguishing feature of atrial flutter is the relatively high voltage, 0.2 to 0.8 mv of the P waves in lead II and sometimes in lead III. Leads II and III record higher voltage than lead I because the circus movement pathway of the impulse causing atrial flutter almost always passes around and around the atria from top to bottom. This means that the current flow in the atria extends up and down the body more or less parallel to the axes of leads II and III, which causes strong voltages in these leads but weak voltages in lead I.

ATRIAL FIBRILLATION

Figure 17–15 illustrates the electrocardiogram during atrial fibrillation. As discussed in Chapter 14, numerous small depolarization waves spread in all directions through the atria during atrial fibrillation. Because the waves are weak and because many of them are of opposite polarity at any given time, they usually almost completely neutralize each other. Therefore, in the electrocardiogram one can see either no P waves from the atria or a fine, high frequency, very low voltage wavy record. On the other hand the QRS-T complexes are completely normal unless there is some pathology of the ventricles, but their timing is very irregular for the following reasons:

Since the fibrillatory depolarization waves in the atria are completely irregular, the intervals between impulses arriving at the A-V node are extremely variable. Therefore, an impulse may arrive at the A-V node immedi-

Figure 17–15. Atrial fibrillation (lead I).

ately after the node itself is out of its refractory period from its previous discharge, or it may not arrive there for several additional tenths of a second. Consequently, the rhythm of the ventricles is very irregular, many of the ventricular beats falling quite close together and many far apart, as illustrated in Figure 17–15. The overall ventricular rate is 125 to 150 beats per minute in most instances.

The pumping effectiveness of the heart in atrial fibrillation is considerably depressed because the ventricles often do not have sufficient time to fill between beats. Therefore, one of the principles of treating patients with atrial fibrillation is to slow the rate of the ventricles. This is ordinarily done by administering digitalis, which causes a prolonged refractory period of the A-V node, probably by stimulating the vagal nerves to the node, thereby decreasing the number of impulses conducted from the atria into the ventricles and preventing any two ventricular beats from occurring close together. Heart function may be greatly improved when the rate is thus depressed to less than 100 beats per minute.

Pulse Deficit in Atrial Fibrillation. In the same manner that premature contractions can cause a pulse deficit in the radial pulse, so can ventricular beats occurring close together in atrial fibrillation also cause a pulse deficit. Therefore, before treatment of atrial fibrillation one frequently finds a pulse deficit, whereas, after slowing the ventricular rate, the pulse deficit often disappears.

VENTRICULAR FIBRILLATION

In ventricular fibrillation the electrocardiogram is extremely bizarre, as seen in Figure 17–16, and ordinarily shows no tendency toward a regular rhythm of any type. In the early phases of ventricular fibrillation relatively large masses of muscle contract simultaneously, and this causes strong though irregular waves in the electrocardiogram. However, after only a few seconds the coarse contractions of the ventricles disappear, and the electrocardiogram changes into a new pattern of low voltage, very irregular waves. Thus, no repetitive electrocardiographic pattern can be ascribed to ventricular fibrillation except that the electrical potentials constantly and spasmodically change because the currents in the heart flow first in one direction, then in another, and rarely repeat any specific cycle.

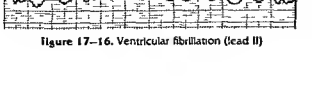


Figure 17–16. Ventricular fibrillation (lead II)

The voltages of the waves in the electrocardiogram in ventricular fibrillation are usually about 0.5 mv when ventricular fibrillation first begins, but these decay rapidly so that after 20 to 30 seconds they are usually only 0.2 to 0.3 mv. Minute voltages of 0.1 mv or less may be recorded for 10 minutes or longer after ventricular fibrillation begins. As already pointed out, ventricular fibrillation is lethal unless it is stopped by some heroic therapy such as immediate electroshock through the heart, as explained in Chapter 14.

REFERENCES

Braun, H A, and Diettert, G A ECG Arrhythmia Interpretation A Programmed Text for Health Care Personnel Reston, Va, Reston Publishing Co, 1979

Brown, H F Electrophysiology of the sinoatrial node *Physiol Rev*, 62 505, 1982

Burch, G E, and Winsor, T A Primer of Electrocardiography 6th Ed Philadelphia, Lea & Febiger, 1972

Chou, T Electrocardiography in Clinical Practice New York, Grune & Stratton, 1979

Chung, E K Electrocardiography Practical Applications With Vectorial Principles, 2nd Ed Hagerstown, Md, Harper & Row, 1980

Cranefield, P F, and Wit, A L Cardiac arrhythmias *Annu. Rev Physiol*, 41 459, 1979

Fletcher, G F Dynamic Electrocardiographic Recording Mt Kisco, N Y, Futura Publishing Co, 1979

Gilmour, R F, Jr, and Zipes, D P Slow inward current and cardiac arrhythmias *Am J Cardiol*, 55 89B 1985

Goldman, M J Principles of Clinical Electrocardiography Los Altos, CA, Lange Medical Publications, 1982

Guyton, A C and Crowell, J W A stereovectorcardiograph *J Lab Clin Med*, 40 726, 1952

Harris, C C A Primer of Cardiac Arrhythmias A Self-Instructional Program St Louis, C V Mosby, 1979

Lapman, B Clinical Electrocardiography Chicago, Year Book Medical Publishers, 1984

McAnulty, J, and Rahimtoola, S Prognosis in bundle branch block *Annu Rev Med* 32 499, 1981

Meijler, F L Atrioventricular conduction versus heart size from mouse to whale *J Am Coll Cardiol*, 3 563 1985

Moe, G K Mechanisms of cardiac dysrhythmias *In* Frohlich, E D (ed) Pathophysiology, 2nd Ed Philadelphia, J B Lippincott Co, 1976, p 83

Morganroth, J Ambulatory Holter electrocardiography Choice of technologies and clinical uses *Ann Intern Med* 102 73, 1985

Moss, A J Prediction and prevention of sudden cardiac death. *Annu Rev Med*, 31 1, 1980

Narula, O S. (ed.) Cardiac Arrhythmias. Electrophysiology, Diagnosis, and Management Baltimore, Williams & Wilkins, 1979

Phibbs, B The Cardiac Arrhythmias St. Louis, C V Mosby, 1978.

Puck, A., and Langendorf, R Interpretation of Complex Arrhythmias. Philadelphia, Lea & Febiger, 1980

Pordy, L. Computer Electrocardiography Present Status and Criteria. Mt. Kisco, N Y., Futura Publishing Co., 1977

Scher, A M, and Spach, M S Cardiac depolarization and repolarization and the electrocardiogram *In* Berne, R. M, *et al* (eds) Handbook of Physiology. Sec 2, Vol I Baltimore, Williams & Wilkins, 1979, p 357

Summerall, C P Monitoring Heart Rhythm. New York, John Wiley & Sons, 1982

V

THE CIRCULATION

18 ■ Physics of Blood, Blood Flow, and Pressure: Hemodynamics

19 ■ The Systemic Circulation

20 ■ Local Control of Blood Flow by the Tissues, and Nervous and Humoral Regulation

21 ■ Arterial Pressure Regulation: I. Rapid Pressure Control by Nervous Reflexes and Other Mechanisms

22 ■ Arterial Pressure Regulation: II. Long-Term Control by the Renal-Body Fluid Mechanism and the Renin-Angiotensin System: Mechanisms of Hypertension

23 ■ Cardiac Output, Venous Return, and Their Regulation

24 ■ The Pulmonary Circulation

25 ■ The Coronary Circulation and Ischemic Heart Disease

26 ■ Cardiac Failure

27 ■ Heart Sounds; Dynamics of Valvular and Congenital Heart Defects

28 ■ Circulatory Shock and Physiology of Its Treatment

29 ■ Muscle Blood Flow During Exercise; Cerebral, Splanchnic, and Skin Blood Flows

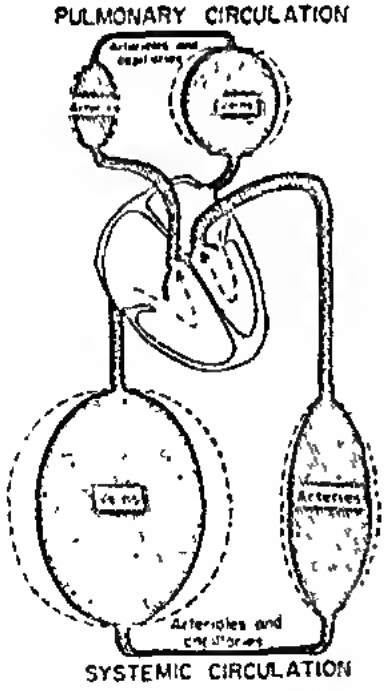
18

Physics of Blood, Blood Flow, and Pressure: Hemodynamics

THE CIRCULATORY SYSTEM AS A "CIRCUIT"

The most important feature of the circulation that must always be kept in mind is that it is a continuous circuit. That is, if a given amount of blood is pumped by the heart, this same amount must also flow through each subdivision of the circulation.

Figure 18–1 illustrates the general plan of the circulation, showing the two major subdivisions, the *systemic circulation* and the *pulmonary circulation*. In the figure the arteries of each subdivisio are represented by a single distensible chamb and all the veins by another even larger distensib chamber, and the arterioles and capillaries repr sent very small connections between the arteri and veins. Blood flows with almost no resistanc in all the large vessels of the circulation, but th is not the case in arterioles and capillaries. T cause blood to flow through these small "resi tance" vessels, the heart pumps blood into th arteries under high pressure—normally at a sy tolic pressure of about 120 mm Hg in the system system and 22 mm Hg in the pulmonary system.



18–1. Representation of the circulation, showing both distensible and the resistive portions of the systemic and pulmonary circulations.

As a first step toward explaining the overa function of the circulation, this chapter will discus the physical characteristics of blood itself and the the physical principles of blood flow through th vessels, especially the interrelationships amon pressure, flow, and resistance. The study of thes interrelationships and other basic physical princ ples of blood circulation is called *hemodynamics*.

THE PHYSICAL CHARACTERISTICS OF BLOOD

Blood is a viscous fluid composed of *cells* an *plasma*. More than 99 per cent of the cells are re blood cells; this means that for practical purpose the white blood cells play almost no role in deter mining the physical characteristics of the blood.

THE HEMATOCRIT

The per cent of the blood that is cells is calle the hematocrit. Thus, if a person has a hematocri of 40, 40 per cent of the blood volume is cells an the remainder is plasma. The hematocrit of norma man averages about 42, while that of norma woman averages about 38. These values vary tre

mendously, depending upon whether or not the person has anemia, the degree of bodily activity, and the altitude at which the person resides. These effects were discussed in relation to the red blood cells and their function in Chapter 4.

Blood hematocrit is determined by centrifuging blood in a calibrated tube such as that shown in Figure 18–2. The calibration allows direct reading of the per cent of cells.

Effect of Hematocrit on Blood Viscosity. Blood is several times as viscous as water, and this viscosity increases the difficulty with which the blood flows through the small vessels. The greater the percentage of cells in the blood—that is, the greater the hematocrit—the more friction there is between successive layers of blood, and this friction determines viscosity. Therefore, the viscosity of blood increases drastically as the hematocrit increases, as illustrated in Figure 18–3. If we consider the viscosity of water to be 1, then the viscosity of whole blood at normal hematocrit is about 3; this means that three times as much pressure is required to force whole water through the same tube. Note that when the hematocrit rises to 60 or 70, which it often does in polycythemia as discussed in Chapter 4, the blood viscosity can become as great as ten times that of water, and its flow through blood vessels is greatly retarded.

Another factor that affects blood viscosity is the concentration and types of proteins in the plasma, but these effects are so much less important than the effect of hematocrit that they are not significant considerations in most hemodynamic studies. The viscosity of blood plasma is about 1.5 times that of water.

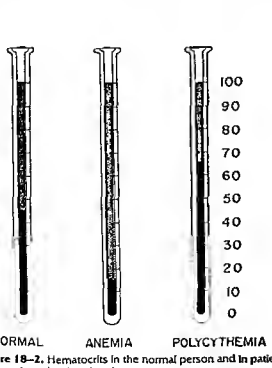


Figure 18–2. Hematocrits in the normal person and in patients with anemia and polycythemia.

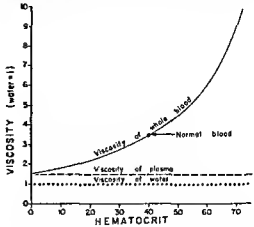


Figure 18–3. Effect of hematocrit on viscosity

Since most resistance in the circulatory system occurs in the very small blood vessels, it is especially important to know how blood viscosity affects blood flow in these minute vessels. At least three additional factors besides hematocrit and plasma proteins affect blood viscosity in these vessels:

1. Blood flow in very minute vessels exhibits far less viscous effect than it does in large vessels. This is called the *Fahraeus-Lindqvist effect.* It begins to appear when the vessel diameter falls below approximately 1.5 mm. In vessels as small as capillaries, this effect is so prominent that the viscosity of whole blood is as little as one half that in large vessels. The Fahraeus-Lindqvist effect is caused by alignment of the red cells as they pass through the vessels. That is, the red cells, instead of moving randomly, line up and move through the vessels as a single plug, thus eliminating the viscous resistance that occurs internally in the blood itself. The Fahraeus-Lindqvist effect, however, is probably more than offset by the following two effects under most conditions.

2. The viscosity of blood increases tremendously as its velocity of flow decreases. Since the velocity of blood flow in the small vessels is extremely minute, often less than 1 mm per second, blood viscosity can increase as much as ten-fold from this factor alone. This effect is partly caused by adherence of the red cells to each other (formation of rouleaux and larger aggregates) and to the vessel walls.

3. Cells also often become stuck at constrictions in small blood vessels; this happens especially in capillaries where the nuclei of endothelial cells protrude into the capillary lumen. When this occurs, blood flow can become totally blocked for a fraction of a second, for several seconds, or for much longer periods of time, thus giving an apparent effect of greatly increased viscosity.

Because of these special effects that occur in the minute vessels of the circulatory system, it has

been impossible to derive an exact mathematical relationship that describes how the hematocrit affects viscosity in the minute vessels, which is the place in the circulatory system where viscosity plays its most important role. Nevertheless, because some of these effects tend to decrease viscosity and others tend to increase viscosity, it is reasonable to assume that the overall viscous effects in the small vessels are approximately equivalent to those that occur in the larger vessels.

PLASMA

Plasma is part of the extracellular fluid of the body. It is almost identical to the interstitial fluid found between the tissue cells except for one major difference: plasma contains about 7 per cent protein, while interstitial fluid contains an average of only 2 per cent protein. The reason for this difference is that plasma protein leaks only slightly through the capillary pores into the interstitial spaces. As a result, most of the plasma protein is held in the circulatory system, and that which does leak is eventually returned to the circulation by the lymph vessels Therefore, the plasma protein concentration is about 3 5 times that of the fluid outside tha capillaries.

The Types of Protein in Plasma. The plasma protein is divided into three major types, as follows:

	Grams Per Cent
Albumin	4.5
Globulins	2.5
Fibrinogen	0.3

The primary function of the *albumin* (and of the other types of protein to a lesser extent) is to cause osmotic pressure at the capillary membrane. This *pressure, called colloid osmotic pressure, prevents* the fluid of the plasma from leaking out of the capillaries into the interstitial spaces. This function is so important that it will be discussed in detail in Chapter 30.

The globulins are divided into three major types: alpha, beta, and gamma globulins. The *alpha* and *beta globulins* perform diverse functions in the circulation, such as transporting various substances by combining reversibly with them, acting as substrate for formation of other substances, and transporting protein itself from one part of the body to another. The *gamma globulins*, and to a lesser extent the *beta globulins*, play a special role in protecting the body against infection, for it is these globulins that are mainly the *antibodies* that resist infection and toxicity, thus providing the [illegible] with what we call *immunity*. The function of [illegible] ity was discussed in detail in Chapter 6.

The *fibrinogen* of plasma is of basic importance [illegible] blood clotting and was discussed in Chapter 8

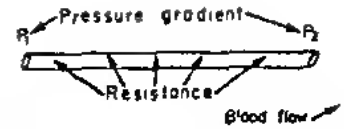


Figure 18–4. Relationships among pressure, resistance, blood flow

INTERRELATIONSHIPS AMONG PRESSURE, FLOW, AND RESISTANCE

Flow through a blood vessel is determined entirely by two factors: (1) the *pressure difference* between the two ends of the vessel, which is the force that pushes the blood through the vessel, and (2) the impediment to blood flow through the vessel, which is called vascular *resistance*. Figure 18–4 illustrates these relationships, showing a blood vessel segment located anywhere in the circulatory system.

P_1 represents the pressure at the origin of the vessel; at the other end the pressure is P_2. The flow through the vessel can be calculated by the following formula, which is called *Ohms law*:

$$Q = \frac{\Delta P}{R} \qquad (1)$$

in which Q is blood flow, ΔP is the pressure difference (P_1-P_2) between the two ends of the vessel, and R is the resistance This formula states, in effect, that the blood flow is directly proportional to the pressure difference but inversely proportional to the resistance.

It should be noted especially that the *difference* in pressure between the two ends of the vessel determines the rate of flow and not the absolute *pressure in the vessel For instance, if the pressure* at both ends of the segment were 100 mm Hg and yet no difference existed between the two ends, there would be no flow despite the presence of 100 mm Hg pressure.

Ohm's law expresses the most important of all the relationships that the student needs to understand to comprehend the hemodynamics of the circulation Because of the extreme importance of this formula, the student should also become familiar with its other two algebraic forms:

$$\Delta P = Q \times R \qquad (2)$$

$$R = \frac{\Delta P}{Q} \qquad (3)$$

BLOOD FLOW

Blood flow means simply the quantity of blood that passes a given point in the circulation in a

given period of time. Ordinarily, blood flow is expressed in *milliliters* or *liters per minute*, but it can be expressed in milliliters per second or in any other unit of flow.

The overall blood flow in the circulation of an adult person at rest is about 5000 ml per minute. This is called the *cardiac output* because it is the amount of blood pumped by the heart in a unit period of time.

Methods for Measuring Blood Flow. Many different mechanical or mechanoelectrical devices can be inserted in series with a blood vessel, or in some instances applied to the outside of the vessel, to measure flow. These are called simply *flowmeters*.

The Electromagnetic Flowmeter. In recent years several new devices have been developed that can be used to measure blood flow in a vessel without opening it. One of the most important of these is the electromagnetic flowmeter, the principles of which are illustrated in Figure 18–5. Figure 18–5A shows generation of electromotive force in a wire that is moved rapidly through a magnetic field. This is the well-known principle for production of electricity by the electric generator. Figure 18–5B shows that exactly the same principle applies for generation of electromotive force in blood when it moves through a magnetic field. In this case, a blood vessel is placed between the poles of a strong magnet, and electrodes are placed on the two sides of the vessel perpendicular to the magnetic lines of force. When blood flows through the vessel, electrical voltage proportional to the rate of flow is generated between the two electrodes, and this is recorded using an appropriate meter or electronic apparatus. Figure 18–5C illustrates an actual "probe" that is placed on a large blood vessel to record its blood flow. This probe contains both the strong magnet and the electrodes.

A special advantage of the electromagnetic flowmeter is that it can record changes in flow that occur in less than 0.01 second, allowing accurate recording of pulsatile changes in flow as well as steady flow.

The Ultrasonic Doppler Flowmeter. Another type of flowmeter that can be applied to the outside of the vessel and that has many of the same advantages as the electromagnetic flowmeter is the ultrasonic Doppler flowmeter illustrated in Figure 18–6. A minute piezoelectric crystal is mounted in the wall of the device. This crystal, when energized with an appropriate electronic apparatus, transmits sound at a frequency of several million cycles per second downstream along the flowing blood. A portion of the sound is reflected by the flowing red blood cells so that reflected sound waves travel backward from the blood toward the crystal. However, these reflected waves have a lower frequency than the transmitted wave because the red cells are moving away from the transmitter crystal. This is called the Doppler effect. (It is the same effect that one experiences when a train approaches a listener and passes by while blowing the whistle. Once the whistle has passed by the person, the pitch of the sound from the whistle suddenly becomes much lower than when the train is approaching.) The transmitted wave is intermittently cut off and the reflected wave is received back onto the crystal, then amplified greatly by the electronic apparatus. Another portion of the apparatus determines the frequency difference between the transmitted wave and the reflected wave, thus also determining the velocity of blood flow.

Like the electromagnetic flowmeter, the ultrasonic Doppler flowmeter is capable of recording very rapid, pulsatile changes in flow as well as steady flow.

Laminar Flow of Blood in Vessels. When blood flows at a steady rate through a long, smooth vessel, it flows in *streamlines*, with each layer of blood remaining the same distance from the wall. Also, the central portion of the blood stays in the center of the vessel. This type of flow is called *laminar flow* or *streamline flow*, and it is opposite to *turbulent flow*, which is blood

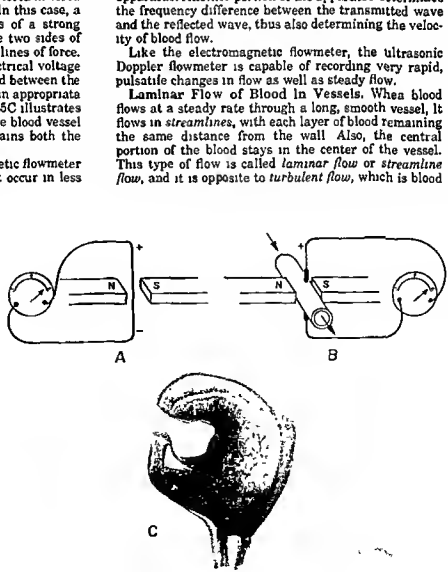


Figure 18–5. A flowmeter of the electromagnetic type, showing A, generation of an electromotive force in a wire as it passes through an electromagnetic field, B, generation of an electromotive force in electrodes on a blood vessel when the vessel is placed in a strong magnetic field and blood flows through the vessel, and C, a modern electromagnetic flowmeter "probe" for chronic implantation around blood vessels.

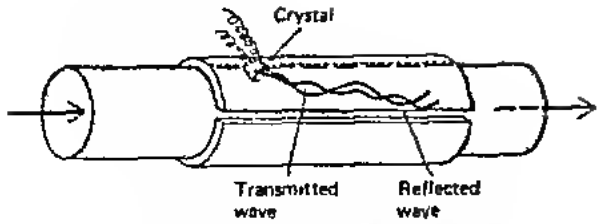


Figure 18–6. An ultrasonic Doppler flowmeter.

flowing in all direction in the vessel and continually mixing within the vessel, as will be discussed below

Parabolic Velocity Profile During Laminar Flow. When laminar flow occurs, the velocity of flow in the center of the vessel is far greater than that toward the outer edges This is illustrated by the experiment shown in Figure 18–7 In vessel A are two different fluids, the one to the left colored by a dye and the one to the right a clear fluid, but there is no flow in the vessel Then the fluids are made to flow, and a parabolic interface develops between the two fluids, as shown 1 second later in vessel B, illustrating that the portion of fluid adjacent to the vessel wall has hardly moved at all, the portion slightly away from the wall has moved a small distance, and the portion in the center of the vessel has moved a long distance This effect is called the parabolic profile for the velocity of blood flow

The cause of the parabolic profile is the following· The fluid molecules touching the wall hardly move because of adherence to the vessel wall The next layer of molecules slips over these, the third layer over the second, the fourth layer over the third, and so forth Therefore, the fluid in the middle of the vessel can move rapidly because many layers of molecules exist between the middle of the vessel and the vessel wall, all of these capable of slipping over each other, while those portions of fluid near the wall do not have this advantage

Turbulent Flow of Blood Under Some Conditions. When the rate of blood flow becomes too great, when it passes an obstruction in a vessel, when it makes a sharp turn, or when it passes over a rough surface, the flow may then become *turbulent* rather than streamline Turbulent flow means that the blood flows crosswise in the vessel as well as along the vessel, usually forming whorls in the blood called *eddy currents* These are similar to the whirlpools that one frequently sees in a rapidly flowing river at a point of obstruction

When eddy currents are present, blood flows with much greater resistance than when the flow is streamline because the eddies add tremendously to the overall friction of flow in the vessel

The tendency for turbulent flow increases in direct proportion to the velocity of blood flow, in direct proportion to the diameter of the blood vessel, and inversely proportional to the viscosity of the blood divided by its density in accordance with the following equation:

$$Re = \frac{v \cdot d}{\frac{\eta}{\rho}} \qquad (4)$$

in which Re is *Reynolds' number* and is the measure of the tendency for turbulence to occur, v is the velocity of blood flow (in centimeters), η is the viscosity (in poises) and ρ is density. The viscosity of blood is normally about 1/30 poise, and the density is only slightly greater than 1 When Reynolds' number rises above 200 to 400, turbulent flow will occur at some branches of vessels but will die out along the smooth portions of the vessels However, when Reynolds' number rises above approximately 2000, turbulence will usually occur even in a straight, smooth vessel Reynolds' number for flow in the vascular system even normally rises to 200 to 400 in large arteries; as a result there is almost always some turbulence of flow at the branches of the vessels In the proximal portions of the aorta and pulmonary artery, Reynolds' number can rise to several thousand during the rapid phase of ejection by the ventricles; this causes considerable turbulence in the proximal aorta and pulmonary artery where many conditions are appropriate for turbulence (1) high velocity of blood flow, (2) pulsatile nature of the flow, (3) sudden change in vessel diameter, and (4) large vessel diameter.

However, in small vessels, Reynolds' number is almost never high enough to cause turbulence.

BLOOD PRESSURE

The Standard Units of Pressure. Blood pressure is almost always measured in *millimeters of mercury (mm Hg)* because the mercury manometer (shown in Figure 18–8) has been used as the standard reference for measuring blood pressure throughout the history of physiology Actually,

Figure 18–7. An experiment illustrating laminar blood flow **A**, two separate fluids before flow begins. *B*, the same fluids 1 second after flow begins.

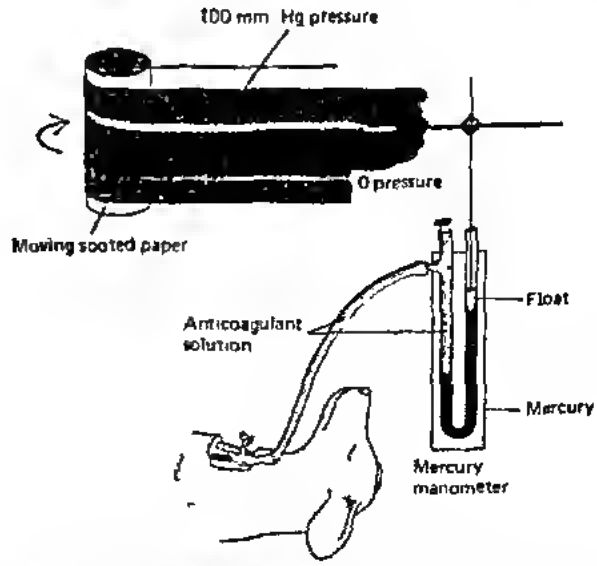


Figure 18–8. Recording arterial pressure with a mercury manometer, a method that has been used in the manner shown above for recording pressure throughout the history of physiology

blood pressure means the *force exerted by the blood against any unit area of the vessel wall.* When one says that the pressure in a vessel is 50 mm Hg, this means that the force exerted is sufficient to push a column of mercury up to a level 50 mm high. If the pressure is 100 mm Hg, it will push the column of mercury up to 100 mm.

Occasionally, pressure is measured in *centimeters of water.* A pressure of 10 cm of water means a pressure sufficient to raise a column of water to a height of 10 cm. *One millimeter of mercury equals 1.36 cm of water* because the specific gravity of mercury is 13.6 times that of water, and 1 cm is 10 times as great as 1 mm. Dividing 13.6 by 10, we derive the factor 1.36.

Measurement of Blood Pressure Using the Mercury Manometer. Figure 18–8 illustrates a standard mercury manometer for measuring blood pressure. A cannula or catheter is inserted into an artery, a vein, or even the heart, and the pressure from the cannula or catheter is transmitted to the left-hand side of the manometer where it pushes the mercury down while raising the right-hand mercury column. The difference between the two levels of mercury is approximately equal to the pressure in the circulation in terms of millimeters of mercury.

High-Fidelity Methods for Measuring Blood Pressure. Unfortunately, the mercury in the mercury manometer has so much *inertia* that it cannot rise and fall rapidly. For this reason the *mercury manometer,* though excellent for recording steady pressures, cannot respond to pressure changes that occur more rapidly than approximately one cycle every 2 to 3 seconds. Whenever it is desired to record rapidly changing pressures, some other type of pressure recorder is needed. Figure 18–9 demonstrates the basic principles of three electronic pressure *transducers* commonly used for converting pressure into electrical signals and then recording the pressure on a high-speed electrical recorder. Each of these transducers employs a very thin and highly stretched metal membrane which forms one wall of the fluid chamber. The fluid chamber in turn is connected through a needle or catheter with the vessel in which the pressure is to be measured. Pressure variations in the vessel cause changes of pressure in the chamber beneath the membrane. When the pressure is high the membrane bulges outward slightly, and when low it returns toward its resting position.

In Figure 18–9A a simple metal plate is placed a few thousandths of an inch above the membrane. When the membrane bulges outward, the *capacitance* between the plate and membrane increases, and this change in capacitance can be recorded by an appropriate electronic system.

In Figure 18–9B a small iron slug rests on the membrane, and this can be displaced upward into a coil. Movement of the iron changes the *inductance* of the coil, and this, too, can be recorded electronically.

Finally, in Figure 18–9C a very thin, stretched resistance wire is connected to the membrane. When this wire is greatly stretched its resistance increases, and when less stretched the resistance decreases. These changes also can be recorded by means of an electronic system.

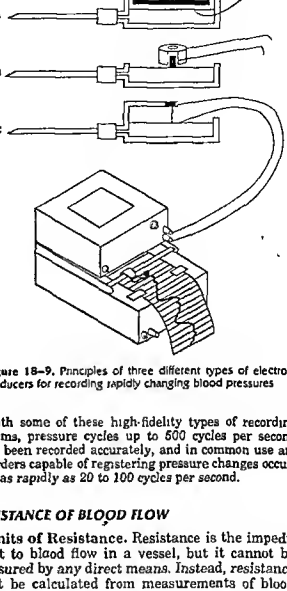


Figure 18–9. Principles of three different types of electronic transducers for recording rapidly changing blood pressures

With some of these high-fidelity types of recording systems, pressure cycles up to 500 cycles per second have been recorded accurately, and in common use are recorders capable of registering pressure changes occurring as rapidly as 20 to 100 cycles per second.

RESISTANCE OF BLOOD FLOW

Units of Resistance. Resistance is the impediment to blood flow in a vessel, but it cannot be measured by any direct means. Instead, resistance must be calculated from measurements of blood flow and pressure difference in the vessel. If the pressure difference between two points in a vessel is 1 mm Hg and the flow is 1 ml/sec, then the resistance is said to be 1 *peripheral resistance unit,* usually abbreviated *PRU.*

Expression of Resistance in CGS Units. Occasionally, a basic physical unit called the CGS (centimeters, grams, seconds) unit is used to express resistance. This unit is *dyne seconds/centimeters.*[8] Resistance in these units can be calculated by the following formula:

$$R\left(in\ \frac{\text{dyne sec}}{\text{cm}^5}\right) = \frac{1333 \times \text{mm Hg}}{\text{ml/sec}} \qquad (5)$$

Total Peripheral Resistance and Total Pulmonary Resistance. The rate of blood flow through the circulatory system when a person is at rest is close to 100 ml/sec, and the pressure difference from the systemic arteries to the systemic veins is about 100 mm Hg. Therefore, in round figures the resistance of the entire s

circulation, called the *total peripheral resistance*, is approximately 100/100 or 1 PRU. In some conditions in which all the blood vessels throughout the body become strongly constricted, the total peripheral resistance rises to as high as 4 PRU, and when the vessels become greatly dilated it can fall to as little as 0.2 PRU.

In the pulmonary system the mean arterial pressure averages 16 mm Hg and the mean left atrial pressure averages 2 mm Hg, giving a net pressure difference of 14 mm Therefore, in round figures the *total pulmonary resistance* at rest calculates to be about 0.14 PRU This can increase in disease conditions to as high as 1 PRU and can fall in certain physiological states, such as exercise, to as low as 0 04 PRU.

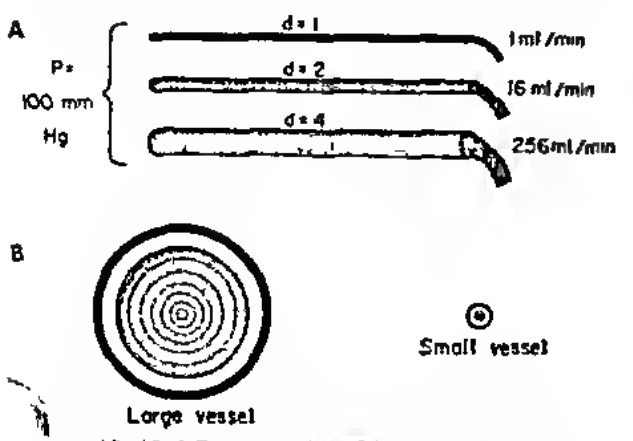
"Conductance" of Blood in a Vessel and Its Relationship to Resistance. Conductance is a measure of the blood flow through a vessel for a given pressure difference. This is generally expressed in terms of ml/sec/mm Hg pressure, but it can also be expressed in terms of liters/sec/mm Hg or in any other units of blood flow and pressure.

It is immediately evident that conductance is the reciprocal of resistance in accord with the following equation:

$$\text{Conductance} = \frac{1}{\text{Resistance}} \qquad (6)$$

Effect of Vascular Diameter on Conductance. Slight changes in the diameter of a vessel cause tremendous changes in its ability to conduct blood when the blood flow is streamline This is illustrated forcefully by the experiment in Figure 18–10A, which shows three separate vessels with relative diameters of 1, 2, and 4 but with the same pressure difference of 100 mm Hg between the two ends of the vessels. Though the diameters of these vessels increase only fourfold, the respective flows are 1, 16, and 256 ml/mm, which is a 256-fold increase in flow. Thus, the conductance of the vessel increases in proportion to the *fourth power of the diameter*, in accord with the following formula:

$$\text{Conductance} \propto \text{Diameter}^4 \qquad (7)$$



18–10. *A*, Demonstration of the effect of vessel diameter on blood flow *B*. Concentric rings of blood flowing at different velocities the farther away from the vessel wall, the faster the flow

Poiseuille's Law. The cause of this great increase in conductance with an increase in diameter can be explained by referring to Figure 18–10B This illustrates cross-sections of a large and a small vessel. The concentric rings inside these vessels indicate that the velocity of flow in each ring is different from that in the other rings because of *laminar* flow, which was discussed earlier in the chapter That is, the blood in the ring touching the wall of the vessel is flowing hardly at all because of its adherence to the vascular endothelium. The next ring of blood slips past the first ring and, therefore, flows at a more rapid velocity. The third, fourth, fifth, and sixth rings likewise flow at progressively increasing velocities. Thus, the blood that is very near the wall of the vessel flows extremely slowly, while that in the middle of the vessel flows extremely rapidly.

In the small vessel essentially all the blood is very near the wall so that the extremely rapidly flowing central stream of blood simply does not exist.

By integrating the velocities of all the concentric rings of flowing blood one can derive the following formula relating *mean velocity* of the total blood flow to vascular radius

$$v = \frac{\Delta P \cdot r^2}{8\,\eta l} \qquad (8)$$

in which v is velocity (in centimeters per second), ΔP is the pressure gradient (in dynes per square centimeter), r is the radius of the vessel (in centimeters), η is the viscosity (in poises), and l is the length of the vessel (in centimeters)

The quantity of blood that will flow through a vessel in a given period of time is equal to the velocity of flow times the cross-sectional area according to the following equation

$$Q = v\pi r^2 \qquad (9)$$

in which Q is the rate of blood flow (in milliliters per second) and πr^2 is the cross-sectional area (in square centimeters).

Now let us substitute the value for velocity of blood flow from Equation 8 into Equation 9. This gives the following equation, which is known as Poiseuille's law:

$$Q = \frac{\pi \Delta P r^4}{8\,\eta l} \qquad (10)$$

Note particularly in this equation that the rate of blood flow is directly proportional to the *fourth power of the radius* of the vessel, which illustrates once again that the diameter of a blood vessel plays by far the greatest role of all factors in determining the rate of blood flow through the vessel.

Summary of the Different Factors That Affect Conductance and Resistance. In the equation representing Poiseuille's law, Q represents flow, and ΔP represents the pressure difference. The remainder of the

equation represents the conductance in accordance with the following equation:

$$C = \frac{\pi r^4}{8\,\eta l} \qquad (11)$$

And, since conductance is the reciprocal of resistance, the following equation shows the factors that affect resistance:

$$R = \frac{8\,\eta l}{\pi r^4} \qquad (12)$$

Thus, note that the resistance of a vessel is directly proportional to the *blood viscosity* (η) and *length* (l) of the vessel but inversely proportional to the *fourth power of the radius* (r^4).

Resistance to Blood Flow Through Series Vessels. In Figure 18–11A are two vessels connected in series, having resistances of R_1 and R_2. It is immediately evident that the total resistance is equal to the sum of the two, or

$$R_{(total)} = R_1 + R_2 \qquad (13)$$

Furthermore, it is equally evident that any number of resistances in series with each other must be added together. For instance, the total peripheral resistance is equal to the resistance of the arteries plus that of the arterioles plus that of the capillaries plus that of the veins.

Resistance of Vessels in Parallel. Shown in Figure 18–11B are four vessels, *connected in parallel*, with respective resistances of R_1, R_2, R_3, and R_4. However, the diameters of the vessels are not exactly the same It is obvious that for a given pressure difference, far greater amounts of blood will flow through this system than through any one of the vessels alone Therefore, the total resistance is far less than the resistance of any single vessel.

To calculate the total resistance in Figure 18–11B one first determines the *conductance* of each of the vessels, which is equal to the reciprocal of the resistance, or $\frac{1}{R_1}$, $\frac{1}{R_2}$, $\frac{1}{R_3}$, and $\frac{1}{R_4}$. The total conductance [C (total)] of all the vessels is equal to the sum of the individual conductances:

$$C_{(total)} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \frac{1}{R_4} \qquad (14)$$

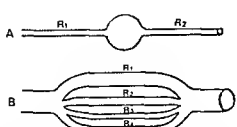


Figure 18–11. Vascular resistances. *A*, in series and *B*, in parallel.

And resistance through the parallel circuit is the reciprocal of the total conductance, or

$$R_{(total)} = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \frac{1}{R_4}} \qquad (15)$$

Effect of Pressure on Vascular Resistance and Tissue Blood Flow

From the discussions thus far, one would expect that an increase in arterial pressure would cause a proportionate increase in blood flow through the various tissues of the body However, the effect of pressure on blood flow is far greater than one would expect, as illustrated in Figure 18–12 The reason for this is that an increase in arterial pressure not only increases the force that tends to push blood through the vessels but also distends the vessels at the same time, which decreases their resistance Thus, increased pressure increases the flow in two different ways Consequently, for most tissues, blood flow at 100 mm Hg arterial pressure is usually about four times as great as the blood flow at 50 mm Hg

Note also in Figure 18–12 the large changes in blood flow that can be caused by either increased or decreased sympathetic stimulation of the peripheral blood vessels Thus, inhibition of sympathetic stimulation greatly dilates the vessels and therefore increases the blood flow, sometimes as much as twofold or more. Conversely, very strong sympathetic stimulation can constrict the vessels so much that blood flow can be decreased sometimes to as low as zero for short periods of time despite high arterial pressure.

Cessation of Blood Flow at Very Low Arterial Pressures. Careful study of the lower ends of the curves in Figure 18–12 will show that the blood flow ceases completely even when a small amount of pressure is still in the arteries There are two major reasons for this

First, as the arterial pressure falls to progressively lower levels, the extent to which the pressure stretches

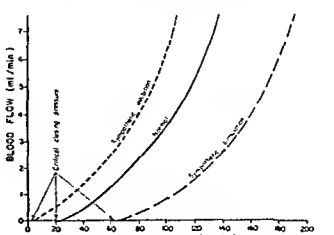


Figure 18–12. Effect of arterial pressure on blood flow through a blood vessel at different degrees of vascular tone caused by increased or decreased sympathetic stimulation of the vessels.

the vessel wall diminishes, finally reaching a point at which the pressure inside the vessel is no longer capable of opposing the muscular tone in the vessel wall and keeping the vessel open This point at which the vessel closes is called the "critical closing pressure" A physical law, the *law of Laplace*, helps explain this closure of the vessels This law states that the circumferential force tending to stretch the muscle fibers in the vessel wall is proportional to the *diameter of the vessel* times the *pressure inside the vessel*, or $F = D \cdot P$. Therefore, as the pressure in the vessel falls there is double reason for the diameter to decrease first, the decrease in pressure causes decreased diameter; then this decreased diameter decreases the force of stretch still more, causing still further decrease in diameter Thus, a vicious circle can develop that will cause vessels to close very suddenly. This is the origin of the term "critical closing pressure"

Second, another cause of the cessation of blood flow at low arterial pressures is the tendency for the blood cells to become stuck in the arterioles and capillaries when the pressure becomes too little to force the cells past obstruction points, especially points where the nuclei of the endothelial cells protrude into the vessel lumens The pressure level at which the red cells will barely begin to flow through a narrow vessel is called the "yield pressure," but the term critical closing pressure is frequently used in an all-inclusive manner to include this yield effect as well To illustrate the importance of the blood cells, the average arterial pressure level at which blood flow ceases in most tissues is about 20 mm Hg when whole blood flows through the tissue and only 5 to 10 mm Hg when plasma without blood cells flows through

VASCULAR DISTENSIBILITY

The diameter of blood vessels, unlike that of metal pipes and glass tubes, increases as the internal pressure increases, because blood vessels are *distensible*. However, the vascular distensibilities differ greatly in different segments of the circulation, and, as we shall see, this affects significantly the operation of the circulatory system under many changing physiological conditions.

Units of Vascular Distensibility. Vascular distensibility is normally expressed as the fractional increase in volume for each millimeter mercury rise in pressure in accordance with the following formula.

$$\text{Vascular distensibility} = \frac{\text{Increase in volume}}{\text{Increase in pressure} \times \text{Original volume}} \quad (16)$$

That is, if 1 mm Hg causes a vessel originally containing 10 ml of blood to increase its volume by 1 ml, then the distensibility would be 0 1 per mm Hg or 10 per cent per mm Hg.

Difference in Distensibility of the Arteries and the Veins. Anatomically, the walls of arteries are far stronger than those of veins Consequently, the veins, on the average, are about six to ten times as distensible as the arteries. That is, a given rise in pressure will cause about six to ten times as much extra blood to fill a vein as an artery of comparable size.

In the pulmonary circulation the veins are very similar to those of the systemic veins. However, the pulmonary arteries, which normally operate under pressures about one sixth those in the systemic arterial system, have distensibilities only about one half those of veins, rather than one eighth as is true of the systemic arteries.

VASCULAR COMPLIANCE (OR CAPACITANCE)

Usually in hemodynamic studies it is much more important to know the *total quantity of blood* that can be stored in a given portion of the circulation for each mm Hg pressure rise than to know the distensibility of the individual vessels. This value is sometimes called the *overall distensibility* or *total distensibility*, or it can be expressed still more precisely by either of the terms *compliance* or *capacitance*, which are physical terms meaning the increase in volume that causes a given increase in pressure in any vascular area as follows:

$$\text{Vascular compliance} = \frac{\text{Increase in volume}}{\text{Increase in pressure}} \quad (17)$$

Compliance and distensibility are quite different A highly distensible vessel that has a very slight volume may have far less compliance than a much less distensible vessel that has a very large volume, for *compliance is equal to distensibility × volume*

The compliance of a vein is about 24 times that of its corresponding artery because it is about 8 times as distensible and it has a volume about 3 times as great ($8 \times 3 = 24$).

VOLUME-PRESSURE CURVES OF THE ARTERIAL AND VENOUS CIRCULATIONS

A convenient method for expressing the relationship of pressure to volume in a vessel or in a large portion of the circulation is the so-called *volume-pressure curve* (also frequently called the *pressure-volume curve*). The two solid curves of Figure 18–13 represent respectively the volume-pressure curves of the normal arterial and venous systems, showing that when the arterial system, including the larger arteries, small arteries, and arterioles, is filled with approximately 750 ml of blood the mean arterial pressure is 100 mm Hg, but when filled with only 500 ml the pressure falls to zero

In the entire venous system, on the other hand, the volume of blood normally is about 2500 ml, and tremendous changes in this volume are required to change the venous pressure only a few millimeters of mercury.

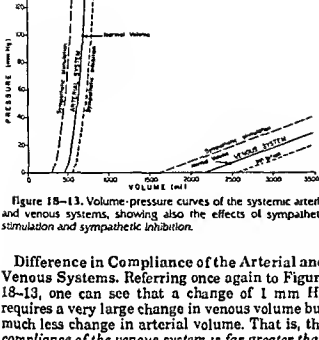


Figure 18–13. Volume-pressure curves of the systemic arterial and venous systems, showing also the effects of sympathetic stimulation and sympathetic inhibition.

Difference in Compliance of the Arterial and Venous Systems. Referring once again to Figure 18–13, one can see that a change of 1 mm Hg requires a very large change in venous volume but much less change in arterial volume. That is, the *compliance of the venous system is far greater than the compliance of the arteries—about 24 times as great.*

This difference in compliance is particularly important because it means that tremendous amounts of blood can be stored in the veins with only slight changes in pressure. Therefore, the veins are frequently called *storage areas* of the circulation.

Effect of Sympathetic Stimulation or Sympathetic Inhibition on the Volume-Pressure Relationships of the Arterial and Venous Systems. Also shown in Figure 18–13 are the volume-pressure curves of the arterial and venous systems during moderate sympathetic stimulation and during sympathetic inhibition. It is evident that sympathetic stimulation, with its concomitant increase in smooth muscle tone in the vascular walls, increases the pressure at each volume of the arteries or veins while, on the other hand, sympathetic inhibition decreases the pressure at each respective volume. Obviously, control of the vessels in this manner by the sympathetics is a valuable means for diminishing the dimensions of one segment of the circulation, thus transferring blood to other segments. For instance, an increase in vascular tone throughout the systemic circulation often causes large volumes of blood to shift into the heart, which is a major way in which pumping by the heart is increased.

Sympathetic control of vascular capacity is also especially important during hemorrhage. Enhancement of the sympathetic tone of the vessels, especially of the veins, reduces the dimensions of the circulatory system, and the circulation continues to operate almost normally even when as much as 25 per cent of the total blood volume has been lost.

"MEAN CIRCULATORY FILLING PRESSURE" AND VOLUME-PRESSURE CURVES OF THE ENTIRE CIRCULATORY SYSTEM

THE MEAN CIRCULATORY FILLING PRESSURE

The mean circulatory filling pressure (also called the "mean circulatory pressure" or the "static pressure") is a measure of the degree of filling of the entire circulatory system. That is, it is the pressure that would be measured in the circulation if one could instantaneously stop all blood flow and bring all the pressures in the circulation immediately to equilibrium. The mean circulatory filling pressure has been estimated reasonably accurately in dogs within 2 to 5 seconds after the heart has been stopped. To do this the heart is thrown into fibrillation by an electrical stimulus, and blood is pumped rapidly from the systemic arteries to the veins to cause pressure equilibration between the two major chambers of the circulation

The mean circulatory filling pressure measured in the above manner in the normal pentobarbital-anesthetized dog is almost exactly 7 mm Hg. It is believed to be about this same value in the human being. However, many different factors can change it, including especially change in blood volume and increased or decreased sympathetic stimulation.

The mean circulatory filling pressure is *one of the major factors that determines the rate at which blood flows from the vascular tree into the right atrium of the heart, which in turn determines the cardiac output.* This is so important that this relationship will be explained in detail in Chapter 23; it will also be discussed in relation to blood volume regulation in Chapter 36.

VOLUME-PRESSURE CURVES OF THE ENTIRE CIRCULATION

Figure 18–14 illustrates the changes in the mean circulatory filling pressure as the total blood volume increases (1) under normal conditions, (2) during strong sympathetic stimulation, and (3) during complete sympathetic inhibition. The point marked by the arrow is the operating point of the normal circulation: a mean circulatory filling pressure of 7 mm Hg and a blood volume of 5000 ml. However, *if blood is lost from the circulatory system,* the mean circulatory filling pressure falls to a lower value. If increased amounts of blood are added, the mean circulatory filling pressure rises accordingly.

The *compliance* of the entire circulatory system in the human being—that is, the increase in volume when the filling pressure increases—is approximately 100 ml for each 1 mm rise in mean circulatory filling pressure.

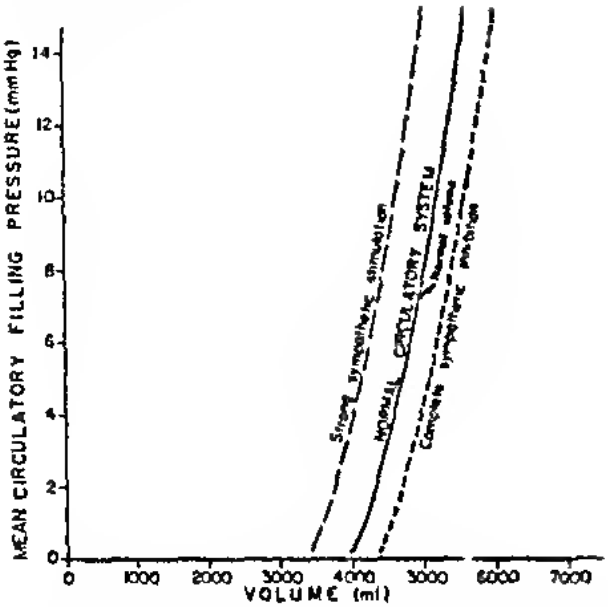


Figure 18–14. Volume-pressure curves of the entire circulation, illustrating the effect of strong sympathetic stimulation and complete sympathetic inhibition.

Sympathetic stimulation and *inhibition* affect the volume-pressure curves of the entire circulatory system in the same way that they affect the volume-pressure curves of the individual parts of the circulation, as illustrated by the two dashed curves of Figure 18–14. That is, for any given blood volume, the mean circulatory filling pressure rises two- to fourfold with strong sympathetic stimulation and falls markedly when the sympathetics are inhibited. This is an extremely important factor in the regulation of blood flow into the heart and thereby for regulating the cardiac output. For instance, during exercise, sympathetic activity increases the mean circulatory filling pressure severalfold, in this way helping to increase the cardiac output.

DELAYED COMPLIANCE (STRESS-RELAXATION) OF VESSELS

The term "delayed compliance" means that a vessel whose pressure is increased by increased volume gradually loses much of this pressure over a period of many minutes because of progressive stretch of the vessel. Likewise, a vessel exposed to constantly increased pressure becomes progressively enlarged.

Figure 18–15 diagrams one of the effects of delayed compliance. In this figure, the pressure is being recorded in a small segment of a vein that is occluded at both ends. Then, an extra volume of blood is suddenly injected into the segment until the pressure rises from 5 to 12 Hg. Even though none of the blood is removed after is injected, the pressure nevertheless begins to fall ly and approaches approximately 9 mm Hg several minutes. In other words, the volume of injected caused immediate *elastic* distension of the vein, but then the smooth muscle fibers of the vein

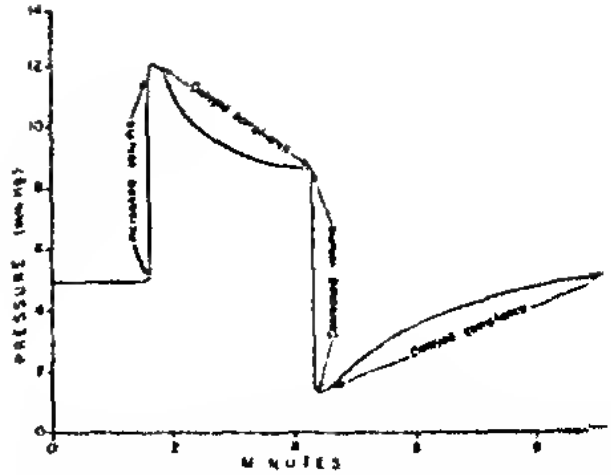


Figure 18–15. Effect on the intravascular pressure of injecting a small volume of blood into a venous segment, illustrating the principles of delayed compliance.

began to "creep" to longer lengths, and their tensions correspondingly decreased. Therefore, the pressure rose markedly at first but then decreased with time. This effect is a characteristic of all smooth muscle tissue and is called *stress relaxation*, which was explained in Chapter 12.

After the delayed increase in compliance had taken place in the experiment illustrated in Figure 18–15, the extra blood volume was suddenly removed, and the pressure immediately fell to a very low value. Subsequently, the smooth muscle fibers began to readjust their tensions back to their initial values, and after a number of minutes the normal vascular pressure of 5 mm Hg returned.

Delayed compliance occurs only slightly in the arteries but to a much greater extent in the veins. As a result, prolonged elevation of venous pressure can often double the blood volume in the venous tree. This is a valuable mechanism by which the circulation can accommodate much extra blood when necessary, such as following too large a transfusion. Also, delayed compliance in the reverse direction is one of the ways in which the circulation automatically adjusts itself over a period of minutes or hours to diminished blood volume after serious hemorrhage.

REFERENCES

American Physiological Society. Peripheral Circulation and Organ Blood Flow. Washington, D.C.: American Physiological Society, 1983.
American Physiological Society. Microcirculation. Washington, D.C.: American Physiological Society, 1984.
Bergel, D. H. (ed.): Cardiovascular Fluid Dynamics. New York, Academic Press, 1972.
Bessis, M., *et al* (eds.): Red Cell Rheology. New York: Springer-Verlag, 1978.
Branch, D.: Red cell deformability and capillary blood flow. *Physiol. Rev.*, *51*: 679, 1971.
Caro, C. G., and Lever, M. J.: Factors influencing arterial wall mass transport. *Biorheology*, *21*: 197, 1984.
Chien, S., Usami, S., and Skalak, R.: Blood flow in small tubes. *In* Renkin, E. M., and Michel, C. C. (eds.): Handbook of Physiology. Sec. 2, Vol. IV. Bethesda, American Physiological Society, 1984, p. 217.
Cokelet, G. R.: Rheology and hemodynamics. *Annu. Rev. Physiol.*, *42*: 311, 1980.
Cokelet, G. R., *et al* (eds.): Erythrocyte Mechanics and Blood Flow. New York, A. R. Liss, 1979.
Dobrin, P. B.: Mechanical properties of arteries. *Physiol. Rev.*, *58*: 397, 1978.

Dobrin, P. B · Vascular mechanics. *In* Shepherd, J T, and Abboud, F M (eds.) Handbook of Physiology Sec 2, Vol III Bethesda, American Physiological Society, 1983, p 65

Fukada, E., *et al*.. Methods to study rheological properties of blood during clotting *Biorheology*. (Suppl 1) 9, 1984

Fung, Y. C. Introduction to biophysical aspects of microcirculation *In* Kaley, G, and Altura, B M (eds) Microcirculation *Vol I* Baltimore, University Park Press 1977, p 253

Fung, Y C: Rheology of blood in microvessels *In* Kaley, G, and Altura, B M (eds) Microcirculation Vol I Baltimore, University Park Press, 1977, p 279

Gaehtgens, P, Meiselman, H J, and Wayland H Erythrocyte flow velocities in mesenteric microvessels of the cat. *Microvasc Res*, 2 151, 1970.

Gow, B S. Circulation correlates. Vascular impedance, resistance, and capacity *In* Dohr, D F, *et al* (eds.) Handbook of Physiology Sec 2, Vol II Baltimore, Williams & Wilkins, 1980, p 353

Green, H D. Circulation Physical principals *In* Glasser O (ed) Medical Physics Chicago, Year Book Medical Publishers 1944

Gross, J F., and Popel, A. (eds) Mathematics of Microcirculation Phenomena. New York, Raven Press, 1980

Guyton, A. C: Arterial Pressure and Hypertension Philadelphia, W B Saunders Co, 1980

Guyton, A. C, and Greganti, F P A physiologic reference point for measuring circulation pressures in the dog—particularly venous pressure *Am J. Physiol*, *185* 137, 1956

Guyton, A C, *et al* · Pressure-volume curves of the entire arterial and venous systems in the living animal *Am J Physiol 184* 253, 1956

Guyton, A C, *et al* Cardiac Output and Its Regulation Philadelphia, W, B Saunders Co, 1973

Huang, C, and Copley, A L (eds.) Biorheology New York, American Institute of Chemical Engineers, 1978

Hwang, M H C, *et al* (eds.) Quantitative Cardiovascular Studies Clinical Research Applications of Engineering Principles Baltimore, University Park Press, 1978

Intaglietta, M, *et al* · Velocity pressure, flow, and elastic properties in microvessels of cat omentum *Am J Physiol*, *221* 922, 1971

James, D. G (ed.)· Circulation of the Blood Baltimore, University Park Press, 1978

Kenner, T (ed.) Cardiovascular System Dynamics Models and Measurements, New York, Plenum Publishing Corp, 1982

Lassen, N A, *et al* Indicator methods for measurement of organ and tissue blood flow *In* Shepherd, J T, and Abboud, F M (eds) Handbook of Physiology Sec 2, Vol III Bethesda, American Physiological Society, 1983, p 21

Lee, J Pressure flow relationships of single vessels and organs *In* Kaley, G, and Altura, B M (eds) Microcirculation Vol I Baltimore, University Park Press, 1977, p 335

Manning, R. D, Jr, *et al* Essential role of mean circulatory filling pressure in salt induced hypertension *Am J Physiol*, *236* R40, 1979

McDonald, D A Blood Flow in Arteries 2nd Ed Baltimore, Williams & Wilkins, 1974

Murphy, R. A Mechanics of vascular smooth muscle *In* Bohr, D F, *et al* (eds) Handbook of Physiology Sec 2, Vol II Baltimore, Williams & Wilkins, 1980, p 325

O Rourke, M F Vascular impedance in studies of arterial and cardiac function *Physiol Rev* 62 570 1982

Patel, D J, *et al* Hemodynamics *Annu Rev Physiol* 36 125, 1974

Pedley T J The Fluid Mechanics of Large Blood Vessels New York, Cambridge University Press, 1979

Roach, M R Biophysical analyses of blood vessel walls and blood flow *Annu Rev Physiol*, *39* 51 1977

Rodkiewcz, C M Arteries and Arterial Blood Flow New York, Springer Verlag, 1983

Rosenblum, W I Viscosity In vitro versus in vivo In Kaley, G, and Altura, B M (eds) Microcirculation Vol I Baltimore University Park Press, 1977, p 225

Rowan Physics and the Circulation Philadelphia, Heyden & Sons Inc, 1981

Schmid Schonbein, H Microrheology of erythrocytes, blood viscosity, and the distribution of blood flow in the microcirculation *Int Rev Physiol*, *9* 1, 1976

Schneck, D J and Vawter, D L (eds) Biofluid Mechanics New York, Plenum Press, 1980

Sigel, B *et al* Ultrasonic detection of red cell aggregation immediately preceding blood clotting *Invest Radiol 19* 458, 1984

Stehbens, W E (ed) Hemodynamics and the Blood Vessel Wall Springfield, Ill, Charles C Thomas 1978

Taylor, M G Hemodynamics *Annu Rev Physiol* 35 87 1973

Zweifach B W and Lipowsky, H H Pressure flow relations in blood and lymph microcirculation *In* Renkin E M, and Michel, C C (eds) Handbook of Physiology Sec 2 Vol IV Bethesda, American Physiological Society, 1984, p 251

19

The Systemic Circulation

The circulation is divided into the *systemic circulation* and the *pulmonary circulation* And, because the systemic circulation supplies all the tissues of the body except the lungs with blood flow, it is also frequently called the *greater circulation* or *peripheral circulation.*

Though the vascular system in each separate tissue of the body has its own special characteristics, some general principles of vascular function nevertheless apply in all parts of the systemic circulation It is the purpose of the present chapter to discuss these general principles.

The Functional Parts of the Systemic Circulation. Before attempting to discuss the details of function in the systemic circulation, it is important to understand the overall role of each of its parts, as follows:

The function of the *arteries* is to transport blood *under high pressure* to the tissues For this reason the arteries have strong vascular walls, and blood flows rapidly in the arteries.

The *arterioles* are the last small branches of the arterial system, and they act as *control valves* through which blood is released into the capillaries The arteriole has a strong muscular wall that is capable of closing the arteriole completely or of allowing it to be dilated severalfold, thus having the capability of vastly altering blood flow to the capillaries.

The function of the *capillaries* is to exchange fluid, nutrients, electrolytes, hormones, and other substances between the blood and the interstitial spaces For this role, the capillary walls are very thin and are permeable to small molecular substances

The *venules* collect blood from the capillaries; they gradually coalesce into progressively larger veins

The *veins* function as conduits for transport of blood from the tissues back to the heart. Since the pressure in the venous system is very low, the venous walls are thin. Even so, they are muscular, and this allows them to contract or expand and thereby act as a reservoir for extra blood, either a small or large amount depending upon the needs of the body.

PHYSICAL CHARACTERISTICS OF THE SYSTEMIC CIRCULATION

Quantities of Blood in the Different Parts of the Circulation. By far the greater amount of the blood in the circulation is contained in the systemic veins. Figure 19–1 shows this, illustrating that approximately 84 per cent of the entire blood volume of the body is in the systemic circulation, with 64 per cent in the veins, 15 per cent in the arteries, and 5 per cent in the capillaries. The heart contains 7 per cent of the blood, and the pulmonary vessels, 9 per cent. Most surprising is the very low blood volume in the capillaries of the systemic circulation, only about 5 per cent of the total Yet, it is here that the most important function of the systemic circulation occurs, diffusion of substances back and forth between the blood and the tissues. This function is so important that it will be discussed in detail in Chapter 30

Cross-Sectional Areas and Velocities of Blood Flow. If all vessels of each type were put

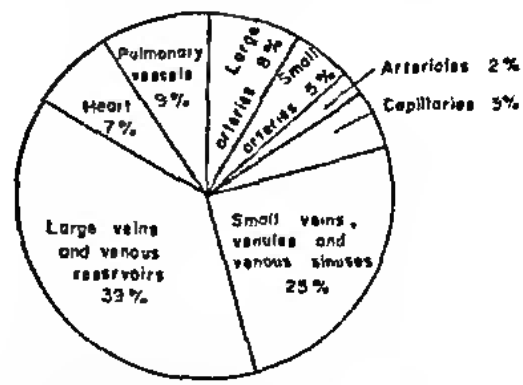


Figure 19–1. Percentage of the total blood volume in each portion of the circulatory system.

side by side, their total cross-sectional areas would be:

	cm^2
Aorta	2.5
Small arteries	20
Arterioles	40
Capillaries	2500
Venules	250
Small veins	80
Venae cavae	8

Note particularly the much larger cross-sectional areas of the veins than of the arteries, averaging about four times those of the corresponding arteries. This explains the very large storage of blood in the venous system in comparison with that in the arterial system.

Because the same volume of blood flows through each segment of the circulation each minute, the velocity of blood flow in each segment is inversely proportional to its cross-sectional area. Thus, under resting conditions, the velocity averages 33 cm per second in the aorta, but 1/1000 of this in the capillaries, or about 0.3 mm per second. However, since the capillaries have a typical length of only 0.3 to 1 mm, the blood remains in the capillaries for only 1 to 3 seconds—a very surprising fact since all diffusion that takes place through the capillary walls must occur in this exceedingly short time.

Pressures and Resistances in the Various Portions of the Systemic Circulation. Because the heart pumps blood continually into the aorta, the pressure in the aorta is obviously high, averaging approximately 100 mm Hg. And, also, because the pumping by the heart is pulsatile, the *arterial pressure fluctuates between a systolic level* of 120 mm Hg and a *diastolic level of* 80 mm Hg, as illustrated in Figure 19–2. As the blood flows through the systemic circulation, its pressure falls progressively to approximately 0 mm Hg by the time it reaches the right atrium.

The decrease in arterial pressure in each segment of the systemic circulation is directly proportional to the vascular resistance in the segment. Thus, in the aorta the resistance is almost zero; therefore, the mean arterial pressure at the end of the aorta is still almost 100 mm Hg. Likewise, the resistance in the large arteries is very slight so that the mean arterial pressure in arteries as small as 3 mm in diameter is still 95 to 97 mm Hg. Then the resistance begins to increase rapidly in the very small arteries, causing the pressure to drop to approximately 85 mm Hg at the beginning of the arterioles.

The resistance of the *arterioles* is greatest of any part of the systemic circulation, accounting for about half the resistance in the entire systemic circulation. Thus, the pressure decreases about 55 mm Hg in the arterioles; therefore, the pressure of the blood as it leaves the arterioles to enter the capillaries is only about 30 mm Hg. Arteriolar resistance is so important to the regulation of blood flow in different tissues of the body that it is discussed in detail later in the chapter and also in the following few chapters, which consider the regulation of the systemic circulation.

The pressure at the arterial ends of the *capillaries* is normally about 30 mm Hg and at the venous ends about 10 mm Hg. Therefore, the pressure decrease in the capillaries is only 20 mm Hg, which illustrates that the capillary resistance is about two fifths that of the arterioles

The pressure at the beginning of the venous system, that is, at the *venules*, is about 10 mm Hg, and this decreases to almost exactly 0 mm Hg at the right atrium. This large decrease in pressure in the veins indicates that the veins have far more *resistance than one would expect for vessels of* their large sizes. Much of this resistance is caused by compression of the veins from the outside, which keeps many of them, especially the venae cavae, collapsed a large share of the time. This effect is discussed later in the chapter

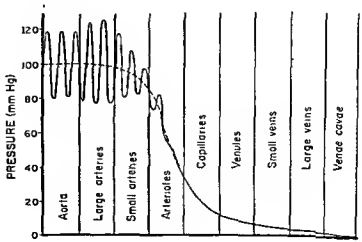


Figure 19–2. Blood pressures in the different portions of the systemic circulatory system.

FUNCTION OF THE LARGE ARTERIES

The large arteries provide two major functions. The first and most obvious of these is to act as conduits to conduct blood to all the peripheral tissues. Under resting conditions, after the blood picks up oxygen in the lungs it generally reaches the tissues in less than 10 seconds; in very heavy exercise, this period is reduced to as little as 2 to 3 seconds.

The second important function of the arteries is to serve as a high pressure reservoir to receive the pulsatile output of blood from the heart and to store some of this blood for part of a second until the next heartbeat. To serve this function, the large arteries are elastic, which means that they stretch as the pulse of blood from the heart enters the arterial tree. Then they recoil to a smoller aize as the blood flows out of the arteries between heartbeats This elastic nature of the arteries is important for two principal reasons: (1) It prevents the pressure from rising extremely high when the blood is pumped into the arterial tree by ventricular contraction. (2) The elasticity maintains a high arterial pressure between heartbeats so that blood can continue to flow through the tissues without interruption Therefore, the normal aortic pressure rises only to 120 mm Hg with each heartbeat and falls only to 80 mm Hg between beats.

THE SMALL ARTERIES, ARTERIOLES, AND CAPILLARIES

Blood flow in each tissue is controlled almost entirely by the degree of contraction or dilatation of the small arteries and arterioles, and it is in the capillaries that the important process of exchange between blood and the interstitial fluid occurs These segments of the circulation are so important that they will receive special discussion in chapters to follow: arterial and arteriolar regulation of blood flow in Chapters 20 and 21 and capillary exchange phenomena in Chapter 30

Upon leaving the small arteries, blood courses through the arterioles, which are only a few millimeters in length and have diameters of from 8 to 50 microns. Each arteriole branches many times to supply 10 to 100 capillaries.

An important characteristic of the small arteries and arterioles is their strong muscular walls. These vessels are intensely innervated by the sympathetic nervous system and, when stimulated, constrict powerfully. Furthermore, many local factors in the tissues—such as the concentrations of oxygen, carbon dioxide, hydrogen ions, and other factors—can directly affect the degree of constriction of the vessels, especially of the arterioles. Therefore, it is mainly the small arteries and arterioles that control the blood flow to the respective tissues, and this flow is normally controlled almost exactly in proportion to the specific ne of each tissue as will be discussed fully in following chapter.

There are approximately 10 billion capillar in the peripheral tissues, and these together h more than 500 square meters of surface area. T thickness of the capillary wall is usually less t 1 micron, and there are small pores in the wall will be explained in Chapter 30) through wh substances can diffuse.

Though the capillary walls are very thin a therefore, very weak, their diameters are also v small. It will be recalled from the discussion the previous chapter of the relationship betw vascular pressure and vascular wall tension t the wall tension is directly proportional to *pressure times the diameter* (law of Laplace). The fore, since the diameter is extremely minute, tension developed in the wall is also extrem minute, which explains why the very thin-wal capillaries can withstand the pressure therein.

EXCHANGE OF FLUID AND NUTRIENTS THROUGH THE CAPILLARY MEMBRANE

Though the detailed dynamics of fluid exchar through the capillary membrane will be presen in Chapter 30, it is important to introduce t subject briefly here.

The most important function of the capillaries is to allow two-way diffusion of nutrients and other substances between the flowing blood and the fluids of the tissues This process allows the blood to carry oxygen to the tissues and to return carbon dioxide from the tissues to the lungs. And such substances as sodium ions, potassium ions, chloride ions, and others diffuse in both directions through the capillary membrane so rapidly that their concentrations are almost exactly the same in both the blood and the tissue fluids

However, one of the major questions that we need to answer is: Why does all the fluid of the blood not leak out of the capillaries into the tissues? The average *functional* capillary pressure (as will be explained in Chapter 30) is about 17 mm Hg, whereas the pressure in the tissue fluid is near zero or even less than zero. Yet, fortunately, the plasma proteins pass hardly at all through the capillary membrane because their molecular sizes are larger than the sizes of the pores in this membrane. These plasma proteins also prevent the leakage of fluid out of the blood into the tissue spaces, which can be explained as follows:

Let us first recall from Chapter 9 the discussion of osmotic *pressure* When one species of molecules in a fluid will not pass through a membrane but all others will, these nonpermeating molecules will cause osmosis of water molecules from the opposite side of the membrane Therefore, because of the high concentration of protein molecules in the plasma, there is a constant tendency for osmosis

of water to occur into the capillaries from the interstitial fluids. The force caused by this osmotic tendency is called the *colloid osmotic pressure* or *osmotic pressure* of the plasma. Under normal circumstances, the tendency for this colloid osmotic pressure to cause movement of fluid into the capillaries is almost exactly as great as the fluid pressure inside the capillaries that is tending to cause movement in the outward direction. Therefore, an almost exact balance is achieved except for a minute amount of leakage of fluid and proteins into the tissues which eventually enters the lymphatic system and is returned to the circulation in that way. Thus, appropriate substances diffuse readily back and forth between the blood and the tissues to provide the nutritive needs of the cells, and yet, at the same time, the volume of the blood is maintained at a normal level. The details of these capillary mechanisms will be discussed in Chapter 30.

THE VEINS AND THEIR FUNCTIONS

For years the veins have been considered to be nothing more than passageways for flow of blood into the heart, but it is rapidly becoming apparent that they perform many functions that are necessary to the operation of the circulation. Especially important, they are capable of constricting and enlarging, of storing large quantities of blood and making this blood available when it is required by the remainder of the circulation, of actually propelling blood forward by means of a so-called "venous pump," and even of helping regulate cardiac output, a function so important that it will be described in detail in Chapter 23.

VENOUS PRESSURES—RIGHT ATRIAL PRESSURE (CENTRAL VENOUS PRESSURE) AND PERIPHERAL PRESSURES

To understand the various functions of the veins, it is first necessary to know something about the pressures in the veins and how they are regulated. Blood from all the systemic veins flows into the right atrium; therefore, the pressure in the right atrium is frequently called the *central venous pressure*. The pressures in the peripheral veins to a great extent depend on the level of this pressure; that is, anything that affects right atrial pressure usually affects venous pressure everywhere in the body.

Right atrial pressure is regulated by a balance between, first, *the ability of the heart to pump blood out of the right atrium* and, second, *the tendency for blood to flow from the peripheral vessels back into the right atrium.*

If the heart is pumping strongly, the right atrial pressure tends to decrease. On the other hand, weakness of the heart tends to elevate the right atrial pressure. Likewise, any effect that causes rapid inflow of blood into the right atrium from the veins tends to elevate the right atrial pressure. Some of the factors that increase this tendency for venous return (and also to increase the right atrial pressure) are (1) increased blood volume, (2) increased large vessel tone throughout the body with resultant increased peripheral venous pressures, and (3) dilatation of the arterioles, which decreases the peripheral resistance and allows rapid flow of blood from the arteries to the veins.

The same factors that regulate right atrial pressure also enter into the regulation of cardiac output, for the amount of blood pumped by the heart depends both on the ability of the heart to pump and the tendency for blood to flow into the heart from the peripheral vessels. Therefore, we will discuss the regulation of right atrial pressure in more depth in Chapter 23 in connection with the regulation of cardiac output.

The *normal right atrial pressure* is approximately 0 mm Hg, which is about equal to the atmospheric pressure around the body. However, it can rise to as high as 20 to 30 mm Hg under very abnormal conditions, such as (a) serious heart failure or (b) following massive transfusion of blood, which will cause excessive quantities of blood to attempt to flow into the heart from the peripheral vessels.

The lower limit to the right atrial pressure is usually about −3 to −5 mm Hg, which is the pressure in the pericardial and intrapleural spaces that surround the heart. The right atrial pressure approaches these very low values when the heart pumps with exceptional vigor or when the flow of blood into the heart from the peripheral vessels is greatly depressed, such as following severe hemorrhage.

Venous Resistance and Peripheral Venous Pressure

Large veins have almost no resistance *when they are distended*. However, as illustrated in Figure 19–3, most of the large veins entering the thorax are compressed at many points by the surrounding tissues so that blood flow is impeded. For instance, the veins from the arms are compressed by their sharp angulation over the first rib. Second, the pressure in the neck veins often falls so low that the atmospheric pressure on the outside of the neck causes them to collapse. Finally, veins coursing through the abdomen are often compressed by different organs and by the intra-abdominal pressure so that usually they are at least partially collapsed to an ovoid or slitlike state. For these reasons the *large veins do usually offer considerable resistance to blood flow*, and because of this the pressure in the peripheral veins is usually 4 to 7 mm Hg greater than the right atrial pressure.

Effect of High Right Atrial Pressure on Pe-

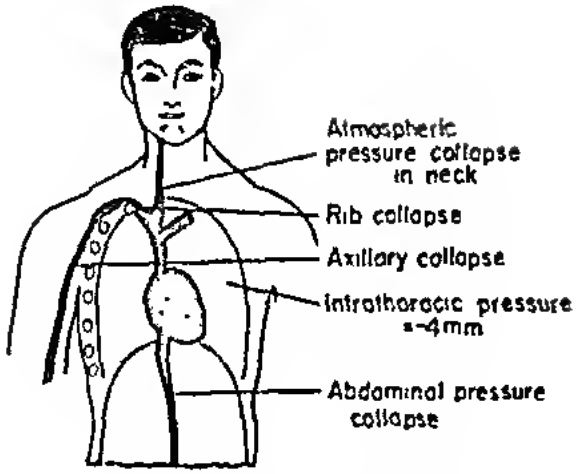


Figure 19–3. Factors tending to collapse the veins entering the thorax.

ripheral Venous Pressure. When the right atrial pressure rises above its normal value of 0 mm Hg, blood begins to back up in the large veins and to open them up. The pressures in the peripheral veins do not rise until all the collapsed points between the peripheral veins and the large veins have opened up. This usually occurs when the right atrial pressure rises to about +4 to 6 mm Hg. If the right atrial pressure then rises still further, the additional increase in pressure is reflected by a corresponding rise in peripheral venous pressure. Since the heart must be greatly weakened to cause a rise in right atrial pressure to as high as 4 to 6 mm Hg, one often finds that the peripheral venous pressure is not elevated in the early stages of heart failure.

Effect of Abdominal Pressure on Venous Pressures of the Leg. The normal pressure in the peritoneal cavity averages about 2 mm Hg, but at times it can rise to as high as 15 to 20 mm Hg as a result of pregnancy, large tumors, or excessive ascites in the peritoneal cavity as explained in Chapter 32. When this happens, the pressure in the veins of the legs must rise *above* the abdominal pressure before the abdominal veins will open and allow the blood to flow from the legs to the heart. Thus, if the intra-abdominal pressure is 20 mm Hg, the lowest possible pressure in the femoral veins is 20 mm Hg.

Effect of "Hydrostatic" Pressure on Venous Pressure

In any body of water, the pressure at the surface of the water is equal to atmospheric pressure, but the pressure rises 1 mm Hg for each 13.6 mm distance below the surface. This pressure results from the weight of the water and therefore is called *hydrostatic pressure*.

Hydrostatic pressure also occurs in the vascular system of the human being because of the weight of the blood in the vessels, as is illustrated in Figure 19–4. When a person is standing, the pressure in the right atrium remains approximately 0 mm Hg because the heart pumps into the arteries any excess blood that attempts to accumulate at this point. However, in an adult *who is standing absolutely still* the pressure in the veins of the feet is approximately +90 mm Hg simply because of the weight of the blood in the veins between the heart and the feet. The venous pressures at other levels of the body lie proportionately between 0 and 90 mm Hg.

In the arm veins, the pressure at the level of the top rib is usually about +6 mm Hg because of compression of the subclavian vein as it passes over this rib. The hydrostatic pressure down the length of the arm is then determined by the distance below the level of this rib. Thus, if the hydrostatic difference between the level of the rib and the hand is 29 mm Hg, this hydrostatic pressure is added to the 6 mm Hg pressure caused by compression of the vein as it crosses the rib, making a total of 35 mm Hg pressure in the veins of the hand.

The neck veins collapse almost completely all the way to the skull owing to atmospheric pressure.

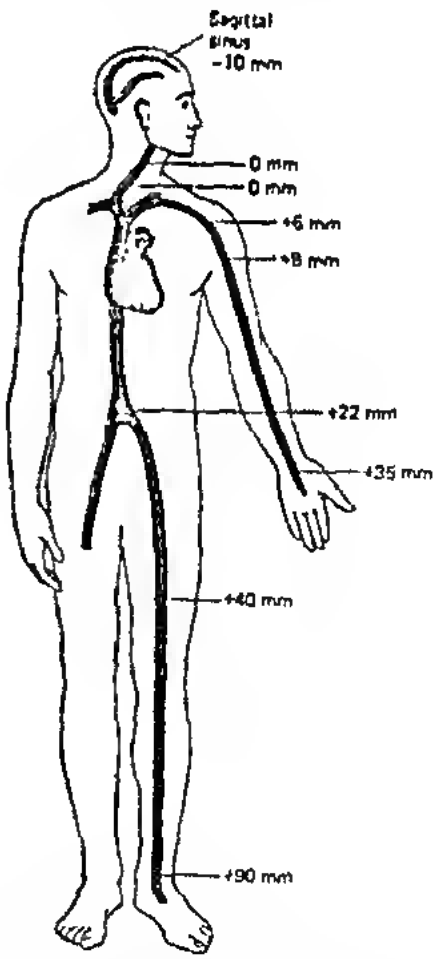


Figure 19–4. Effect of hydrostatic pressure on the venous pressures throughout the body.

on the outside of the neck. This collapse causes the pressure in these veins to remain zero along their entire extent. The reason for this is that any tendency for the pressure to rise above this level opens the veins and allows the pressure to fall back to zero, and any tendency for the pressure to fall below this level collapses the veins still more, which increases their resistance and again returns the pressure back to zero.

The veins inside the skull, however, are in a noncollapsible chamber, and they will not collapse. Consequently, *negative pressure can exist in the dural sinuses of the head;* in the standing position the venous pressure in the sagittal sinus is approximately −10 mm Hg because of the hydrostatic "suction" between the top of the skull and the base of the skull. Therefore, if the sagittal sinus is opened during surgery, air can be sucked immediately into this vein; and the air may even pass downward to cause air embolism in the heart so that the heart valves will not function satisfactorily, and death often ensues.

Effect of the Hydrostatic Factor on Arterial and Other Pressures. The hydrostatic factor also affects the peripheral pressures in the arteries and capillaries as well as in the veins. For instance, a standing person who has an arterial pressure of 100 mm Hg at the level of the heart has an arterial pressure in the feet of about 190 mm Hg. Therefore, any time one states that the arterial pressure is 100 mm Hg, it generally means that this is the pressure at the hydrostatic level of the heart.

Venous Valves, the "Venous Pump," and Venous Pressure

Because of hydrostatic pressure, the venous pressure in the feet would always be about +90 mm Hg in a standing adult were it not for the valves in the veins. However, every time one moves the legs one tightens the muscles and compresses the veins either in the muscles or adjacent to them, and this squeezes the blood out of the veins. Yet, the valves in the veins, as illustrated in Figure 19–5, are arranged so that the direction of blood flow can be only toward the heart. Consequently, every time a person moves the legs or even tenses the muscles, a certain amount of blood is propelled toward the heart, and the pressure in the dependent veins of the body is lowered. This pumping system is known as the "venous pump" or "muscle pump," and it is efficient enough that under ordinary circumstances the venous pressure in the feet of a walking adult remains less than 25 mm Hg.

If the human being stands perfectly still, the venous pump does not work, and the venous pressures in the lower part of the leg can rise to the full hydrostatic value of 90 mm Hg in about 30 seconds. Under such circumstances the pressures within the capillaries also increase greatly, and fluid leaks from the circulatory system into the

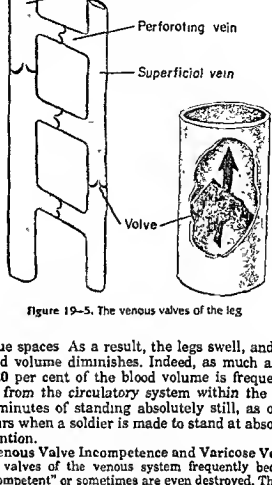


Figure 19–5. The venous valves of the leg

tissue spaces. As a result, the legs swell, and the blood volume diminishes. Indeed, as much as 15 to 20 per cent of the blood volume is frequently lost from the circulatory system within the first 15 minutes of standing absolutely still, as often occurs when a soldier is made to stand at absolute attention.

Venous Valve Incompetence and Varicose Veins. The valves of the venous system frequently become "incompetent" or sometimes are even destroyed. This is especially true when the veins have been overstretched by an excess of venous pressure for a prolonged period of time, as occurs in pregnancy or when one stands most of the time. Stretching the veins obviously increases their cross-sectional areas, but the valves do not increase in size. Therefore, the vanes of the valves will no longer close completely and block reverse blood flow in the enlarged veins. When this develops, the pressure in the veins of the legs increases still more owing to failure of the venous pump; this further increases the size of the veins and finally destroys the function of the valves entirely. Thus, the person develops "varicose veins," which are characterized by large, bulbous protrusions of the veins beneath the skin of the entire leg and particularly of the lower leg. The venous and capillary pressures become very high because of the incompetent venous pump, and leakage of fluid from the capillary blood into the tissues causes constant edema in the legs of these persons whenever they stand for more than a few minutes. The edema in turn prevents adequate diffusion of nutritional materials from the capillaries to the muscle and skin cells so that the muscles become painful and weak, and the skin frequently becomes gangrenous and ulcerates. Obviously, the best treatment for such a condition is continual elevation of the legs to a level at least as high as the heart, but tight binders on the legs are also of considerable aid in preventing the edema and its sequelae.

Clinical Estimation of Venous Pressure. The venous pressure can often be estimated by simply observ-

ing the degree of distension of the peripheral veins—especially the neck veins. For instance, in the sitting position, the neck veins are never distended in the normal person. However, when the right atrial pressure becomes increased to as much as 10 mm Hg, the lower veins of the neck begin to protrude even when one is sitting; when the right atrial pressure rises to as high as 15 mm Hg, essentially all the veins in the neck become distended, and the venous pulse can then usually be seen in the walls of the protruding veins.

Rough estimates of the venous pressure can also be made by raising or lowering an arm of a reclining person while observing the degree of distension of the antecubital or hand veins. As the arm is progressively raised, the veins suddenly collapse, and the level at which they collapse, when referred to the level of the heart, is a rough measure of the peripheral venous pressure.

Direct Measurement of Venous Pressure and Right Atrial Pressure. Venous pressure can be measured with ease by inserting a syringe needle connected to a pressure recorder or to a water manometer directly into a vein.

The only means by which *right atrial pressure* can be measured accurately is by inserting a catheter through the veins into the right atrium. This catheter can then be connected to an appropriate pressure measuring apparatus.

Pressure Reference Level for Measuring Venous and Other Circulatory Pressures

In discussions up to this point we have often spoken of right atrial pressure as being 0 mm Hg and arterial pressure as being 100 mm Hg, but we have not stated the hydrostatic level in the circulatory system to which this pressure is referred. There is one point in the circulatory system at which hydrostatic pressure factors caused by changes in body position usually do not affect the pressure measurement by more than 1 mm Hg. This is at the level of the tricuspid valve, as shown by the crossed axes in Figure 19–6. Therefore, all pressure measurements discussed in this text are referred to the level of the tricuspid valve, which is called the *reference level for pressure measurement*.

The reason for lack of hydrostatic effects at the tricuspid valve is that the heart automatically prevents significant hydrostatic changes in pressure at this point in the following way:

If the pressure at the tricuspid valve rises slightly above normal, then the right ventricle fills to a greater extent than usual, causing the heart to pump blood more rapidly than usual and therefore to decrease the pressure at the tricuspid valve back toward the normal mean value. On the other hand, if the pressure at this point falls, the right ventricle fails to fill adequately, its pumping decreases, and blood dams up in the venous system until the tricuspid pressure again rises to normal value. In other words, *the heart acts as a feedback regulator of pressure* at the tricuspid valve.

The three axes of the reference level, as extrapolated from animals to the human being, are (1) *approximately 61 per cent of the thickness of the chest in the front of the back*, (2) *from side to side almost exactly in the midline*, and (3) *approximately one quarter of the distance above the lower end of the sternum*. A person can be standing, lying on his back, lying on his stomach, lying on either side, or even in the head-down position, and his central venous pressure referred to this reference level will remain almost exactly constant regardless of the position of the body. This does not mean that the pressure at the tricuspid valve is always zero. It may be as low as −4 mm Hg or as high as +20 mm Hg, but, whatever its value, changing the position of the body does not alter it more than 1 mm Hg.

In making arterial pressure measurements, the precise hydrostatic level to which pressures are referred matters much less than for venous pressure measurements because percentagewise a hydrostatic error of as much as 10 or more centimeters (equivalent to 7.4 mm Hg error) still does not affect the arterial pressure reading to a great extent. However, in venous pressure measurements the reference level must be very exact if the measurements are to be significant, because venous pressure changes as small as 1 mm Hg can often result in changes in cardiac output as great as 50 to 100 per cent.

BLOOD RESERVOIR FUNCTION OF THE VEINS

It was pointed out earlier in this chapter that over 60 per cent of all the blood in the circulatory

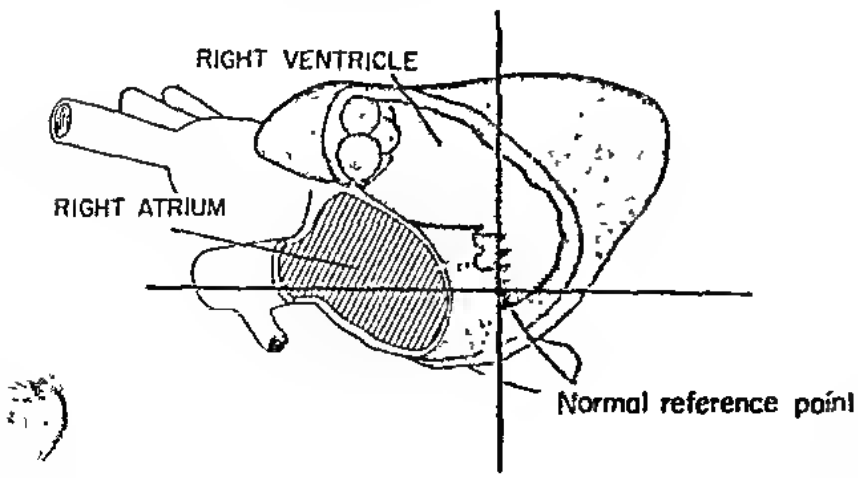


Figure 19–6. Location of the reference point for pressure measurement at the tricuspid valve.

system is in the systemic veins. For this reason it is frequently said that the systemic veins act as a blood reservoir for the circulation. Also, relatively large quantities of blood are present in the veins of the lungs so that these, too, are considered to be a blood reservoir.

When blood is lost from the body and the arterial pressure begins to fall, pressure reflexes are elicited from the carotid sinuses and other pressure-sensitive areas of the circulation, as will be discussed in Chapter 21; these in turn send sympathetic nerve signals to the veins, causing them to constrict, and this takes up much of the slack in the circulatory system caused by the lost blood. Indeed, even after as much as 20 per cent of the total blood volume has been lost, the circulatory system often functions almost normally because of this variable reservoir system of the veins.

Specific Blood Reservoirs. Certain portions of the circulatory system are so extensive that they are specifically called "blood reservoirs." These include (1) the spleen, which can sometimes decrease in size sufficiently to release as much as 150 ml of blood into other areas of the circulation, (2) the liver, the sinuses of which can release several hundred milliliters of blood into the remainder of the circulation, (3) the large abdominal veins, which can contribute as much as 300 ml, and (4) the venous plexus beneath the skin, which can also contribute several hundred milliliters. The heart itself and the lungs, though not parts of the systemic venous reservoir system, must also be considered to be blood reservoirs. The heart, for instance, becomes reduced in size during sympathetic stimulation and in this way can contribute some 50 to 100 ml of blood, and the lungs can contribute another 100 to 200 ml when the pulmonary pressures fall to low values.

Assessment of Filling of the Blood Reservoirs by Measuring Mean Circulatory Filling Pressure

The degree to which the blood reservoirs of the circulation are filled determines to a great extent how well the circulatory system will function. If the blood reservoirs remain overly filled for days or weeks, a person usually develops high blood pressure. Conversely, if they are underfilled, not enough blood will be available in the peripheral circulation to flow back to the heart, and therefore the person will develop circulatory shock.

In experimental animals, one can determine the degree to which the reservoirs are filled with a high degree of accuracy by measuring the mean circulatory filling pressure—that is, by stopping the circulation temporarily and measuring the pressure in it after all pressures have come to equilibrium. As pointed out in the previous chapter, the two factors that determine the mean circulatory filling pressure are (1) the blood volume, and (2) the compliance of the entire circulation.

Because most of the blood volume is in the veins and because the veins are very compliant, the mean circulatory filling pressure is determined mainly by the degree of filling of the venous reservoirs.

The normal mean circulatory filling pressure is about 7 mm Hg, but following massive transfusion this value can rise to as high as 30 to 40 mm Hg, and in congestive heart failure, in which the circulatory volume also becomes greatly increased, the mean circulatory filling pressure can rise to as high as 25 mm Hg.

PRESSURE PULSES IN THE ARTERIES

Since the heart is a pulsatile pump, blood enters the arteries intermittently with each heart beat, causing *pressure pulses* in the arterial system. In the normal young adult the pressure at the height of a pulse, the *systolic pressure*, is about 120 mm Hg and at its lowest point, the *diastolic pressure*, is about 80 mm Hg. The difference between these two pressures, about 40 mm Hg, is called the *pulse pressure*.

Figure 19–7 illustrates a typical, idealized *pressure pulse curve* recorded in the ascending aorta of a human being, showing a very rapid rise in arterial pressure during ventricular systole, followed by a maintained high level of pressure for 0.2 to 0.3 second. This is terminated by a sharp *incisura* or *notch* at the end of systole, followed by a slow decline of pressure back to the diastolic level. The incisura occurs immediately before the aortic valve closes and is caused as follows. When the ventricle relaxes, the intraventricular pressure begins to fall rapidly, and backflow of blood from the aorta into the ventricle allows the aortic pressure also to begin falling. However, the backflow suddenly snaps the aortic valve closed. The momentum that has built up in the backflowing blood brings still more blood into the root of the aorta even after the valve has closed, and this raises the pressure again, thus giving the short positive pressure wave in the record immediately after the incisura.

FACTORS THAT AFFECT THE PULSE PRESSURE

Effect of Stroke Volume and Arterial Compliance. Two major factors affect the pulse pressure: (1) the *stroke volume output* of the heart and (2) the *compliance (total distensibility)* of the arterial tree. A third, less important factor is the character of ejection from the heart during systole.

In general, the greater the stroke volume output, the greater is the amount of blood that must be accommodated in the arterial tree with each heartbeat and, therefore, the greater is the pressure rise and fall during

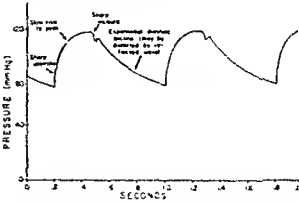


Figure 19–7. A normal pressure pulse contour recorded from the ascending aorta. (From Opdyke, Fed Proc., 11:734, 1952.)

 P P ∝ Stroke Volume

systole and diastole, thus causing a greater pulse pressure

On the other hand, the greater the compliance of the arterial system, the less will be the rise in pressure for a given stroke volume of blood pumped into the arteries. In effect, then, the pulse pressure is determined approximately by the *ratio of stroke volume output to compliance of the arterial tree* Therefore, any condition of the circulation that affects either of these two factors will also affect the pulse pressure.

Factors That Affect the Pulse Pressure by Changing the Stroke Volume Output. So many different circulatory conditions change the stroke volume output that only a few of these can be mentioned.

An *increase in heart rate* while the cardiac output remains constant causes the stroke volume output to decrease in inverse proportion to the increased rate. Therefore, an increase in heart rate decreases the pulse pressure.

A *decrease in total peripheral resistance* allows rapid flow of blood from the arteries to the veins This increases the venous return to the heart and increases the stroke volume output. Therefore, the pulse pressure also greatly increases

An *increase in mean circulatory filling pressure*, if all other circulatory factors remain constant, increases the rate of venous return to the heart and consequently increases both the stroke volume output and the pulse pressure For instance, this is the effect that occurs when a rapid transfusion is given to a person, increasing both the blood volume and the mean circulatory filling pressure

The *character of ejection from the heart* affects the pulse pressure in two ways First, if the duration of systole is long, a large portion of the stroke volume output runs off through the systemic circulation while it is being ejected into the aorta; therefore, the magnitude of the stroke volume effect on pulse pressure is decreased Second, rapid onset of heart contraction sometimes occurs when a heart is beating vigorously, and this causes the pressure in the aorta to rise very high before the blood can run off through the peripheral circulation Therefore, sudden ejection causes a greater pulse pressure than does more prolonged ejection

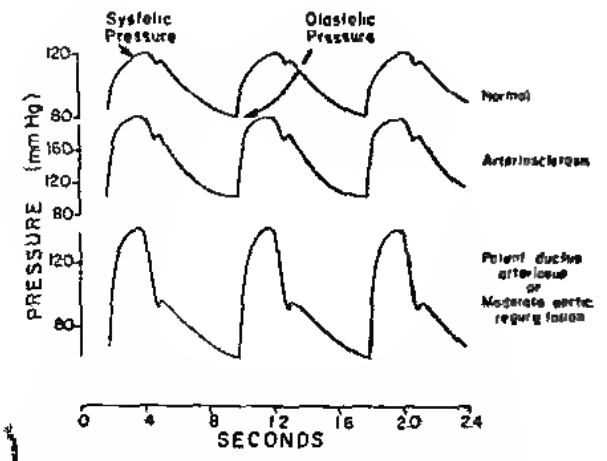


Figure 19–8. Pressure pulse contours in arteriosclerosis, patent ductus arteriosus, and moderate aortic regurgitation.

Decreased Arterial Compliance in Old Age—Arteriosclerosis. In contrast to the many different conditions that can change the stroke volume output, only one major factor often alters the arterial compliance: this is *pathological changes that affect the distensibility of the arterial walls.* This occurs mainly in old age when the arterial walls lose much of their elastic and muscular tissues and are replaced by fibrous tissue and sometimes even calcified plaques that cannot stretch a significant amount This is the well-known condition called *arteriosclerosis* It occurs at a much earlier age in some people than others, but it eventually occurs in all if they live long enough

The vascular changes in arteriosclerosis greatly decrease the compliance of the arterial system, which in turn causes the arterial pressure to rise very high during systole and to fall greatly during diastole as blood runs off from the arteries to the veins. The middle curve of Figure 19–8 illustrates the very great pulse pressure that occurs in arteriosclerosis

ABNORMAL PRESSURE PULSE CONTOURS

Some conditions of the circulation also cause abnormal contours to the pressure pulse wave in addition to altering the pulse pressure Especially distinctive among these are patent ductus arteriosus and aortic regurgitation

Patent Ductus Arteriosus. This is an abnormal condition in which the ductus arteriosus, which carries blood from the pulmonary artery to the aorta during fetal life, fails to close after birth. Instead, the blood now flows backward from the aorta through the open ductus into the pulmonary artery, allowing very rapid runoff of blood from the arterial tree after each heartbeat and a greatly decreased diastolic pressure. However, this is compensated for by a greater than normal stroke volume output because the blood flowing through the ductus passes rapidly through the lungs a second time and re-enters the left heart as extra inflowing blood Therefore, the left heart beats with extra force, and the systolic pressure rises much higher than normal These effects give the pressure pulse contour shown by the lowest curve in Figure 19–8. Here one finds an elevated systolic pressure, a greatly depressed diastolic pressure, a greatly increased pulse pressure, and the incisura occurring at a very low point on the downslope because an excessively large amount of blood runs off from the arteries even before systole is over.

Aortic Regurgitation. Aortic regurgitation means backward flow of blood through the aortic valve because of failure of the valve to close This usually results from heart diseases that destroy the valve, especially rheumatic fever. In aortic regurgitation much of the blood that is pumped into the aorta during systole flows back into the left ventricle during diastole, thus giving a low diastolic pressure. However, this backflow overfills the ventricle before its next beat, therefore, the ventricle pumps a much greater than normal stroke volume output during systole, which causes a high systolic pressure. Thus, the pressure pulse contour is very similar to that of patent ductus arteriosus, as illustrated by the lower curve of Figure 19–8, but is not identical, for in aortic regurgitation the valve sometimes fails to close even partially. When this is true the incisura is entirely absent.

TRANSMISSION OF THE PRESSURE PULSE TO THE PERIPHERY

When the heart ejects blood into the aorta during systole, only the proximal portion of the aorta becomes distended at first, and it is only in this portion of the arterial tree that the pressure rises immediately. The cause of this is the inertia of the blood in the aorta, which prevents its sudden movement away from the rentral arteries into the peripheral arteries.

However, the rising pressure in the central aorta rapidly overcomes the inertia of the blood, causing the rising pressure to move farther and farther out into the arterial tree. Figure 19–9 illustrates this *transmission of the pressure pulse* down the aorta, the more distal portions of the aorta beroming distended as the pressure wave moves forward.

The velocity of transmission of the pressuro pulse along the normal aorta is 3 to 5 meters per second, along the large arterial branches 7 to 10 meters per second, and the smaller arteries 15 to 35 meters per second. In general, the greater the compliance of each vascular segment, the slower is the velocity of transmission, which explains the slowness of transmission in the aorta versus the rapidity of transmission in the far less compliant, small, distal arteries

Note especially that *the velocity of transmission of the pressure pulse is much greater than the velocity of blood flow.* During transmission of the pressuro pulse, only a small amount of blood entering the proximal aorta pushes the more distal blood forward to elevate the pressure in the distal arteries. Therefore, the actual blood ejected by the heart may have traveled only a few centimeters by the time the pressure wave has already reached the distal ends of the arteries In the aorta, the velocity of the pressure pulse is approximately 15 times that of blood flow, while in the more distal arteries the velocity of the pressure pulse may be as great as 100 times the velocity of blood flow.

Augmentation of the Pulse Pressure in the Peripheral Arteries. An interesting phenomenon that often occurs in transmission of pressure *pulses* to the periphery is an increase in pulse pressure. This effect is illustrated in Figure 19–10, which shows considerable enhancement of the pulse pressure more peripherally This effect is caused primarily by *reflection* of the pulse

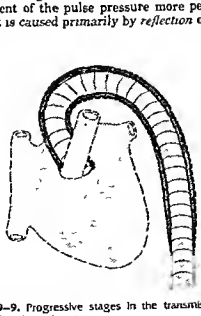


Figure 19–9. Progressive stages in the transmission of the pressure pulse along the aorta.

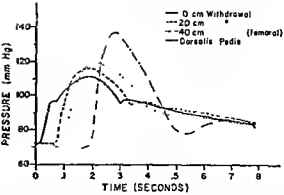


Figure 19–10. Pressure pulse contours in different segments of the arterial tree in humans, showing (1) the delay in the pressure pulse as it spreads more and more peripherally and (2) augmentation of the pulse pressure, especially in the dorsalis pedis artery (Modified from Remington and Wood J Appl Physiol, 9 440, 1956)

wave from peripheral arteries, and it may be explained as follows

When a pressure pulse enters the peripheral arteries and distends them, the rapidly decreasing compliance of the smaller arteries makes it difficult for the pressure pulse to travel any farther. Instead, the surge of pressure in the narrow arteries causes the pulse wave to begin traveling backward along the same vessels from which it had come, thus reflecting the wave backward toward the heart This is analogous to a wave traveling in a bowl of water until it hits the edge On striking the edge, the wave turns around and travels back onto the surface of the water If the returning wave strikes an oncoming wave, the two "summate," causing a much higher wave than would otherwise occur. Such is the case in the arterial tree The first portion of the pressure wave is reflected before the latter portion of the same wave reaches the peripheral arteries. Therefore, the first portion summates with the letter portion, causing a much higher pressure than would otherwise be recorded.

Though this peripheral augmentation of the pulse pressure is of little significance to the function of the circulation, this phenomenon must be remembered whenever pressure measurements are made in peripheral arteries, for systolic pressure is sometimes as much as 20 to 30 per cent above that in the central aorta, and diastolic pressure is often reduced as much as 10 to 15 per cent

Damping of the Pressure Pulse in the Small Arteries and Arterioles. The pressure pulse becomes less and less intense as it passes through the small arteries and arterioles, until it becomes almost absent in the capillaries. This was illustrated in Figure 19–2 and is called *damping* of the pressure pulse.

Damping of the pressure pulse is caused mainly by a *combined effect of vascular distensibility and vascular resistance.* That is, for a pressure wave to travel from one area of an artery to another area, a small amount of blood must flow between the two areas. The resistance in the small arteries and arterioles is great enough that this small flow of blood, and consequently the transmission of pressure, is greatly impeded. At the same time, the distensibility of the small arteries is great : that the small amount of blood that is caused during each pressure pulse produces progressive pressure rise and fall the more distal the pulse proceeds in the vessels.

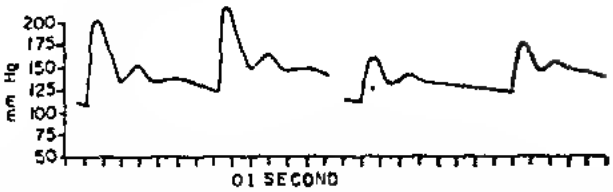


Figure 19–11. Pressure pulse contour in the dorsalis pedis artery of a dog recorded first, under normal conditions and, second during stimulation of the sympathetic nerves supplying the femoral artery

Figure 19–11 illustrates the extreme effect of resistance on damping The two pulses to the left demonstrate a normal pressure pulse contour recorded from the dorsalis pedis artery in a dog To the right is the pressure pulse contour recorded during contraction of the femoral artery and its branches The damping effect of the increased resistance is readily apparent. It is important for the clinician to remember this effect when feeling the peripheral pulses in patients with occlusive vascular disease.

Capillary Pulsation. Pressure pulses are not completely damped by the time they reach the capillaries—the pulse pressure usually still averages 1 mm Hg or more. And under some conditions very severe capillary pulsation results Two major factors can cause abnormal capillary pulsation (1) It occurs when the central pressure pulse is greatly exacerbated, as when the *heart rate is very slow* or when the stroke volume output is greatly increased as a result of *aortic regurgitation, patent ductus arteriosus,* or *even extreme increase in venous return* (2) Another cause of capillary pulsation is *extreme dilatation of the small arteries and arterioles,* which reduces the resistance of these vessels and thereby reduces the damping

Abnormal capillary pulsation can be readily demonstrated by bending the anterior portion of the fingernail downward so that the blood is pressed out of the anterior capillaries of the nail bed. This causes blanching distally while the proximal half of the nail remains red because of blood still in the capillaries. If significant pulsation is occurring in the capillaries, the border between the red and white areas will shift forward as more capillaries become filled during the high pressure phase and will shift backward during the low pressure phase.

THE RADIAL PULSE

Clinically, it has been the habit for many years for a physician to feel the radial pulse of each patient This is done to determine the rate of the heartbeat or, frequently, because of the psychic contact that it gives the doctor with the patient Under certain circumstances, however, the character of the pulse can also be of value in the diagnosis of circulatory diseases

Weak Pulse. A weak pulse at the radial artery usually indicates either (1) greatly decreased central pulse pressure, such as occurs when the stroke volume output is low, or (2) increased damping of the pulse wave caused by vascular spasm, the latter occurs when the sympathetic nervous system becomes overly active, instance, following blood loss or when a person is ng a chill

Pulsus Paradoxus. Occasionally the strength of the lse becomes strong, then weak, then strong, occurring in synchrony with the phases of respiration This is caused by alternate increase and decrease in cardiac output with each respiration. During inspiration, all the blood vessels of the lungs increase in size because of increased negative pressure in the thorax. Therefore, blood collects in the lungs, and the stroke volume output and pulse strength decrease During expiration, opposite effects occur. This is a normal phenomenon in all persons, but it becomes extremely distinct in some conditions, such as in very deep breathing or in cardiac tamponade (compression of the heart from the outside by fluid in the pericardial sac or by a constricted pericardium)

Pulse Deficit. The rhythm of the heart is very irregular in atrial fibrillation or in the case of ectopic beats In these arrhythmias, which were discussed in Chapter 17, two beats of the heart often come so close together that the second beat pumps no blood or very little blood because the left ventricle has too little time to fill between the beats In this circumstance, one can hear the second beat of the heart with a stethoscope applied directly over the heart but cannot feel a pulsation in the radial artery, an effect called a *pulse deficit.* The greater the pulse deficit each minute, the more serious, ordinarily, is the arrhythmia

Pulsus Alternans. In a few conditions the heart beats strongly with one beat and then weakly with the next, causing alternation in the strength of the pulse as well, a condition called *pulsus alternans.* This alternation is frequently caused by enlarged ventricles, the stroke volume output alternating in amount from one heartbeat to the next.

Some instances of pulsus alternans result from abnormal rhythms of the heart, but others perhaps result from a mechanical oscillation that occurs in the circulation as follows A strong heartbeat forces a large amount of blood to the periphery, leaving a deficit of blood in the central vessels Therefore, the next heartbeat is weak But by the third heartbeat, the large surge of blood to the periphery now returns to the input of the heart and a strong heartbeat ensues once again.

PRESSURE PULSES IN THE VEINS

The pressure pulses in the arteries are almost completely damped out before they pass through the capillaries into the systemic veins. However, pulsations are sometimes transmitted backward from the heart to cause pressure pulses in the large veins. This backward transmission of pulses does not occur to a significant extent in the normal circulation because most of the veins, where they lead into the thoracic cavity, are compressed by surrounding tissues, this causes sufficient resistance to damp out the pulsations before they can be transmitted backward into the peripheral veins However, whenever the right atrial pressure is high, especially in heart failure, the veins are then well filled with blood and can easily transmit the pulsations In fact, one can frequently make a diagnosis of heart failure by observing obvious strong pulsations in protruding veins of the neck

It is the *atrial pulsations,* which were discussed in Chapter 13, that are transmitted backward in the distended veins Briefly, these pulsations consist of three separate waves during each heart cycle: (1) the *a wave,* caused by contraction of the atrium, (2) the *c wave,* caused by contraction of the ventricle, and (3) the *v wave,* caused by continued inflow of blood into the atrium

when the A-V valves are closed during ventricular systole.

In severe cardiac failure when the venous pressure is high, the venous pressure waves usually become prominent enough that one can see the veins of the patient's neck pulsating. In earlier days these pulsations were recorded for diagnostic purposes. The *a-c interval* in the recording is approximately equal to the P-Q interval of the electrocardiogram, and before the days of electrocardiography this was often used as a measure of the delay between atrial excitation and ventricular excitation. These waves, therefore, were then of considerable value in diagnosing different degrees of heart block.

REFERENCES

American Physiological Society Peripheral Circulation and Organ Blood Flow. Washington, D C, American Physiological Society, 1983

American Physiological Society Microcirculation Washington, D C, American Physiological Society, 1984

Attinger, E. O Analysis of pulsatile blood flow *Adv Biomed Eng Med Phys*, *1*.1, 1968

D'Agrosa, L. S.. Patterns of venous vasomotion in the bat wing *Am J Physiol*, *218* 530, 1970

Donald, D. E ' Splanchnic circulation *In* Shepherd, J T, and Abboud, F. M. (eds) Handbook of Physiology Sec 2, Vol III Bethesda, American Physiological Society, 1983, p 219

Folkow, B.: Role of the nervous system in the control of vascular tone *Circulation*, *21* 706, 1960

Goodman, A. H., *et al* A television method for measuring capillary red cell velocities *J Appl Physiol*, *37* 126, 1974

Guyton, A. C ' The venous system and its role in the circulation. *Mod Conc Cardiov Dis*, *27* 483, 1958

Guyton, A C ' Peripheral circulation *Annu Rev Physiol*, *21* 239, 1959

Guyton, A. C : Cardiac output and regional circulation *In* Gordon, B L (ed) Clinical Cardiopulmonary Physiology New York, Crune & Stratton, 1969, pp 25–38

Guyton, A. C ' Arterial Pressure and Hypertension Philadelphia, W B Saunders Co, 1980

Guyton, A C, and Jones, C E Central venous pressure Physiological significance and clinical implications *Am Heart J*, *88* 431, 1973

Guyton, A C., *et al*. Evidence for tissue oxygen demand as the major factor causing autoregulation *Circ Res*, *14* 60, 1964

Guyton, A. C, *et al* Cardiac Output and Its Regulation Philadelphia, W B Saunders Co, 1973

Herd, J A. Overall regulation of the circulation. *Annu Rev Physiol*, *32* 289, 1970.

James, D G. (ed) Circulation of the Blood Baltimore, University Park Press, 1978

Mellander, S. Systemic circulation local control *Annu. Rev Physiol*, *32*.313, 1970

Neil, E Peripheral circulation. Historical aspects. *In* Shepherd, J. T, and Abbond, F. M (eds) Handbook of Physiology Sec. 2, Vol, III Bethesda, American Physiological Society, 1983, p. 1

O'Rourke, M F.: Vascular impedance in studies of arterial and cardiac function *Physiol Rev*, *62* 570, 1982.

Pedley, T J ' The Fluid Mechanics of Large Blood Vessels. New York, Cambridge University Press, 1979

Randall, W C (ed.) Nervous Control of Cardiovascular Function. New York, Oxford University Press, 1984

Rodkiewicz, C M Arteries and Arterial Blood Flow New York, Springer-Verlag, 1983

Rothe, C F. Reflex control of veins and vascular capacitance *Physiol Rev*, *63* 1281, 1983

Rothe, C F. Venous system Physiology of the capacitance vessels. *In* Shepherd, J T, and Abboud, F M (eds) Handbook of Physiology Sec 2, Vol III Bethesda, American Physiological Society, 1983, p 397

Rowan Physics and the Circulation. Philadelphia, Heyden & Sons Inc, 1981

Saunders, J B deC M ' The history of venous valves *In* Dickinson, C J, and Marks, J (eds) Developments in Cardiovascular Medicine Lancaster, England, MTP Press, 1978, p 335

Schmidt-Nielsen, K, and Pennycuik, P Capillary density in mammals in relation to body size and oxygen consumption *Am J Physiol*, *200* 746, 1961

Schneck, D J, and Vawter, D. L. (eds) Biofluid Mechanics New York, Plenum Press, 1980

Shepherd, J T, and Vanhoutte, P M The Human Cardiovascular System Facts and Concepts. New York, Raven Press, 1979

Stainsby, W N Autoregulation of blood flow in skeletal muscle during increased metabolic activity *Am J Physiol*, *202* 273, 1962

Stainsby, W N, and Renkin, E M Autoregulation of blood flow in resting skeletal muscle *Am J Physiol*, *201* 117, 1961

Stehbens, W E (ed) Hemodynamics and the Blood Vessel Wall Springfield, Ill, Charles C Thomas, 1978

Uchida, E, and Bohr, D F Myogenic tone in isolated perfused resistance vessels from rats *Am J Physiol*, *216* 1343, 1969

Uchida, E, and Bohr, D F Myogenic tone in isolated perfused vessels Occurrence among vascular beds and along vascular trees *Circ Res*, *25* 549, 1969

Vanhoutte, P M, and Leusen, I (eds.) Mechanics of Vasodilatation New York, S. Karger, 1978

Vanhoutte, P M, *et al* ' Local modulation of the adrenergic neuroeffector interaction in the blood vessel wall *Physiol Rev*, *61*.151, 1981

Wiedeman, M P Dimensions of blood vessels from distributing artery to collecting vein. *Circ. Res*, *12*.375, 1963

Wolf, S, and Werthessen, N T (eds) Dynamics of Arterial Flow. New York, Plenum Press, 1979

Wolf, S., *et al* (eds) Structure and Function of the Circulation. New York, Plenum Press, 1979

Wood, E The Veins. Boston, Little, Brown, 1965

Zeigler, M G Postural hypotension *Annu Rev Med*, *31* 239, 1980

20

Local Control of Blood Flow by the Tissues, and Nervous and Humoral Regulation

The circulatory system is provided with a complex system for control of blood flow to the different parts of the body. In general, the controls are of three major types:

1. Local control of blood flow in each individual tissue, the flow being controlled mainly in proportion to that tissue's need for blood perfusion
2. Nervous control of blood flow, which often affects blood flow in large segments of the systemic circulation, such as shifting blood flow from the nonmuscular vascular beds to the muscles during exercise or changing the blood flow in the skin to control body temperature
3. Humoral control, in which various substances dissolved in the blood such as hormones, ions, or other chemicals can cause either local increase or decrease in tissue flow or widespread generalized changes in flow

LOCAL CONTROL OF BLOOD FLOW IN RESPONSE TO TISSUE NEED

One of the most fundamental and important characteristics of the circulation is the ability of each tissue to control its own local blood flow in proportion to its need. Furthermore, as the need for blood flow changes, the flow follows the change.

However, what are some of the specific needs of the tissues for blood flow? The answer to this is manyfold, including the following:

(1) Delivery of oxygen to the tissues

(2) Delivery of other nutrients like glucose, amino acids, fatty acids, and so forth.

(3) Removal of carbon dioxide from the tissues.

(4) Removal of hydrogen ions from the tissues.

(5) Maintenance of proper concentrations other ions in the tissues.

(6) Transport of various hormones and oth specific substances to the different tissues.

Also, certain organs have special requiremen For instance, blood flow to the skin determin heat loss from the body and in this way hel control the body temperature. And delivery adequate quantities of blood plasma to the kidne allows the kidneys to excrete the waste produ of the body.

We shall see later that some of these facto exert extreme degrees of control over local blo flow.

Variations in Blood Flow in Different Ti sues and Organs. In general, the greater t degree of metabolism in an organ, the greater blood flow. Note, for instance, in Table 20–1 t

Table 20–1 BLOOD FLOW TO DIFFERENT ORGANS AND TISSUES UNDER BASAL CONDITIONS

	Per cent	Ml/min	Ml/ml 100 g
Brain	14	700	50
Heart	4	200	70
Bronchi	2	100	25
Kidneys	22	1100	360
Liver	27	1350	95
Portal	(21)	(1050)	
Arterial	(6)	(300)	
Muscle (inactive state)	15	750	4
Bone	5	250	3
Skin (cool weather)	6	300	3
Thyroid gland	1	50	160
Adrenal glands	0 5	25	300
Other tissues	3 5	175	1.
Total	100 0	5000	—

Based mainly on data compiled by Dr. L. A. Sapirstein.

very large blood flows in the various glandular organs—for example, several hundred ml/min per 100 grams of thyroid or adrenal gland tissue and a blood flow of 95 ml/min per 100 grams of liver.

Also note the extremely large blood flow through the kidneys, 360 ml/min per 100 grams. This extreme amount of flow is required for the kidneys to perform their function of cleansing the blood of waste products.

On the other hand, most surprising is the low blood flow to the resting muscles of the body, even though they constitute between 30 and 40 per cent of the total body mass. However, in the resting state the metabolic activity of the muscles is very low, and so also is the blood flow, only 4 ml/min per 100 grams. Yet, during very heavy exercise, the metabolic activity can increase as much as 50-fold and the blood flow as much as 20- to 25-fold, rising to as high as 80 ml/min per 100 grams.

The Importance of Blood Flow Control by the Local Tissues. One might ask the simple question: Why not simply allow a very large blood flow through every tissue of the body, always enough to supply the tissue's needs whether the activity of the tissue is very little or very great? The answer to this is equally simple: To do this would require many times the amount of blood flow that the heart can pump.

Experiments have shown that the blood flow to each tissue is usually regulated at the minimal level that will supply its requirements, no more, no less. For instance, in tissues for which the most important requirement is delivery of oxygen, the blood flow is only slightly more than the amount required to maintain the tissues at full oxygenation, but no more than this. By controlling the local blood flow in such an exact way, the tissues never suffer from nutritional deficiency, and yet the workload on the heart is kept at a minimum.

MECHANISMS OF BLOOD FLOW CONTROL

Local blood flow control can be divided into two different phases: (1) acute control, and (2) long-term control. Acute control means rapid changes in local blood flow control, occurring within seconds to minutes, to provide a rapid means for maintaining appropriate local tissue conditions. Long-term control, on the other hand, means slow changes in flow over a period of days, weeks, or even months. In general, the long-term changes come about as a result of an increase or decrease in the sizes and numbers of actual blood vessels supplying the tissues. However, before explaining the acute and long-term mechanisms, let us discuss briefly the functional anatomy of the systemic microcirculation that is the basis of the local blood flow control mechanisms.

Functional Anatomy of the Systemic Microcirculation. Though the student will recall that each tissue has its own characteristic vascular system, Figure 20–1 presents a typical local vascular bed; this one is in the connective tissue of the mesentery, which is very easily studied and its components easily analyzed. This figure shows that blood enters a capillary bed through a small *arteriole* and leaves by way of a small *venule*. From the arteriole the blood usually divides and flows through several *metarterioles* before entering the *capillaries*. Some of the capillaries are very large, and they course almost directly to the venule. These are called *preferential channels*. However, most of the capillaries, called the *true capillaries*, branch mainly from the metarterioles and then finally terminate in a venule.

The arterioles have a strong muscular coat, and the metarterioles are surrounded by sparse but highly active smooth muscle fibers. In addition, in many tissues, at each point at which a capillary leaves a metarteriola, a small muscular *precapillary sphincter*, consisting of a single spiraling smooth muscle fiber, surrounds the origin of the capillary. In those tissues that do not have precapillary sphincters, such as in skeletal muscle, the metarterioles serve the same constrictor function for controlling blood flow into the capillaries.

The venules also have a smooth muscle coat, but one that is much less extensive than that of the arterioles.

As will be discussed in more detail later in the chapter, the arterioles especially, and the venules to a much lesser extent, are supplied by extensive innervation from the sympathetic nervous system, and the degree of contraction of these structures is strongly influenced by the intensity of sympathetic signals transmitted from the central nervous system to the blood vessels.

On the other hand, innervation of the metarterioles and the precapillary sphincters is usually very sparse, or even absent in most instances. Instead, the muscle fibers of these two structures are controlled almost entirely by the local factors in the tissues, that is, by the concentrations of oxygen, carbon dioxide, hydrogen

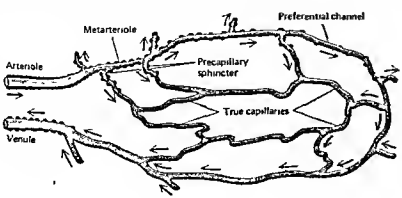


Figure 20–1. Overall structure of a l bed. (From Zweifach: Factors Regulating sure. New York, Josiah Macy, Jr.

theory could explain local blood flow regulation in response to the metabolic needs of the tissues. Perhaps the truth lies in a combination of the two mechanisms.

Possible Role of Other Nutrients Besides Oxygen in the Control of Local Blood Flow. Under special conditions, it has been shown that lack of glucose in the perfusing blood for longer than a few minutes can cause local tissue vasodilation. It is possible that this same effect occurs when other nutrients are also deficient. Indeed, at least one clinical nutritional deficiency leads to marked peripheral vasodilation. This is the vitamin deficiency disease *beriberi*, in which the patient usually has deficiencies of thiamine, niacin, and riboflavin. In this disease the peripheral vascular blood flow increases as much as two- to threefold. Since these vitamins are all concerned with the oxidative phosphorylation mechanism for generating ATP in the local tissues, one could suspect that deficiency of these vitamins leads to diminished smooth muscle contractile ability and therefore leads to the local vasodilation.

Special Examples of "Metabolic" Control of Local Blood Flow

The mechanisms that we have described thus far for local blood flow control are frequently called "metabolic mechanisms" because all of them function in response to the metabolic needs of the tissues. Two additional special instances of metabolic control of local blood flow are *reactive hyperemia* and *active hyperemia*.

Reactive Hyperemia. When the blood supply to a tissue is blocked for a few seconds to several hours and then is unblocked, the flow through the tissue usually increases to about five times normal, the increased flow will continue for a few seconds if the block has lasted a few seconds and sometimes for as long as many hours if the blood flow has been stopped for an hour or more. This phenomenon is called *reactive hyperemia*. Reactive hyperemia is almost certainly another manifestation of the local "metabolic" blood flow regulation mechanism, that is, lack of flow sets into motion all of those factors that cause vasodilatation. Following short periods of vascular occlusion, the extra blood flow during the reactive hyperemia phase lasts long enough to repay almost exactly the tissue oxygen deficit that has accrued during the period of occlusion. This mechanism emphasizes the close connection between local blood flow regulation and delivery of nutrients to the tissues.

Active Hyperemia. When any tissue becomes highly active, such as a muscle during exercise, a gastrointestinal gland during a hypersecretory period, or even the brain during rapid mental activity, the rate of blood flow through the tissue increases. Here again, by simply applying the basic principles of local blood flow control, one can easily understand this so-called *active hyperemia*. The increase in local metabolism causes the cells to devour the tissue fluid nutrients extremely rapidly and also to release large quantities of vasodilator substances. The result obviously would be to dilate the local blood vessels, and therefore, to increase local blood flow. In this way, the active tissue will receive the additional nutrients required to sustain its new level of function.

As pointed out earlier, active hyperemia in skeletal muscle can increase the local blood flow as much as 20- to 25-fold during intense exercise.

"Autoregulation" of Blood Flow When the Arterial Pressure Changes from Normal—"Metabolic" Versus "Myogenic" Mechanisms

In most tissues of the body, an acute increase or a decrease in arterial pressure will cause an immediate increase or decrease in blood flow. However, within less than a minute, the blood flow usually returns most of the way back toward the normal level. Therefore, if a few minutes are allowed for the local blood flow control mechanisms to function properly, the blood flow will be related to arterial pressure approximately in accord with the solid curve labeled "acute" in Figure 20–5. Note that between an arterial pressure of approximately 70 mm Hg and 175 mm Hg the blood flow remains within ± 10 to 25 per cent of the normal value. This maintenance of the flow at nearly normal levels despite marked changes in arterial pressure is called *autoregulation of blood flow*. (The dashed curve of Figure 20–5 illustrates the long-term effects of arterial pressure on blood flow when the pressure changes slowly over a period of weeks or months, as will be discussed in more detail later in the chapter.) For almost a century, two different views have been proposed to explain the acute autoregulation mechanism. These have been called (1) the metabolic theory, and (2) the myogenic theory.

The *metabolic theory* can be understood very easily by simply applying the basic principles of local blood flow regulation already discussed in earlier sections. Thus, when the arterial pressure becomes too great, the excess flow provides too many nutrients to the tissues and also flushes vasodilator substances out of the tissues; both these effects will then cause the blood vessels to constrict. Therefore, the increased pressure will not increase flow greatly because the compensatory constriction will mainly nullify the effect of the pressure.

The *myogenic theory*, on the other hand, suggests that still another mechanism besides those related to tissue

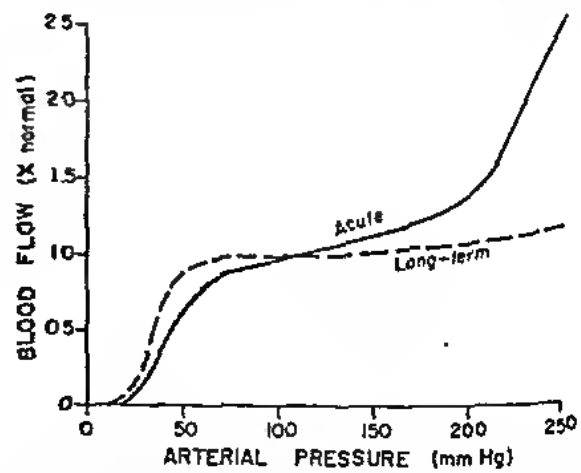


Figure 20–5. Effect of increasing arterial pressure on blood flow through a muscle. The solid curve shows the effect if the arterial pressure is raised over a period of a few minutes. The dashed curve shows the effect if the arterial pressure is raised extremely slowly over a period of many weeks.

metabolism either totally or partially explains the phenomenon of autoregulation. This theory is based on the observation that sudden stretch of small blood vessels will cause the smooth muscle of the vessel wall to contract. Therefore, it is believed that when high arterial pressure stretches the vessel, this in turn causes vascular constriction and reduces the blood flow back near normal. Conversely, at low pressures, the degree of stretch of the vessel is less, so that the smooth muscle relaxes and allows increased flow.

Though some experiments are difficult to explain without invoking the myogenic theory of autoregulation, it is almost certain that myogenic autoregulation is not an extremely powerful functional mechanism in most of the body for a very clear reason: a strong myogenic mechanism everywhere in the body would rapidly lead to death in the following way: An increase in pressure would stretch the blood vessel, which would then cause vasoconstriction. The increased constriction would increase the peripheral resistance and raise the blood pressure still more. This secondary increase in pressure would cause another cycle of stretch followed by still more vasoconstriction and then still more increase in pressure. Thus, a vicious circle would ensue, and if this should occur in all the body at once, the arterial pressure would rise suddenly to such a high level that either the heart would fail or the blood vessels would burst. Even so, in certain isolated tissues, it is difficult to explain some of the phenomena observed without invoking the myogenic theory.

It has been suggested especially that the myogenic mechanism protects the capillaries from excessively high blood pressures. That is, if the pressure in the small arteries and arterioles rises too high, these vessels simply constrict and prevent this high pressure from being transmitted into the capillaries

Special Mechanisms of Local Blood Flow Regulation in Certain Tissues

In certain tissues, blood flow is mainly regulated by other factors besides the requirement for nutrients This is true particularly in the kidney and in the brain, as follows:

"Autoregulation" in the Kidneys. The rate of blood flow through the kidneys remains even more constant than in most other tissues despite marked changes in arterial pressure. Yet the availability of oxygen to the kidneys can be changed tremendously without changing renal blood flow significantly. Therefore, the "autoregulation" of renal blood flow when the arterial pressure changes is undoubtedly different from the usual type of local blood flow regulation.

Two factors that change renal blood flow to a great extent are sodium chloride concentration in the blood and the concentration of end products of protein metabolism; an increase in either of these increases renal blood flow. Recent experiments suggest that it is the concentrations of some of these substances in the renal tubules that control renal blood flow. It is believed that the substances exert their feedback control of afferent arteriolar constriction at the *juxtaglomerular apparatus* where the distal tubule comes in contact with the afferent arteriole. That is, when the concentration of one or more substances becomes too great at this point in the distal tubule, constriction of the afferent arteriole reduces glomerular filtration and thereby causes less entry of the substances into the tubular system. This mechanism will be discussed in detail in Chapter 35 in relation to the physiology of the kidney.

Local Regulation of Cerebral Blood Flow in Response to Tissue CO_2 and Hydrogen Ion Concentration. In the brain, a powerful mechanism for local regulation of blood flow is based on carbon dioxide and hydrogen ion concentrations. When the concentrations of these substances increase, the blood vessels dilate, allowing more rapid blood flow to remove much of the carbon dioxide from the tissues, thereby reducing the CO_2 back toward a normal mean value. This also causes the hydrogen ion concentration to return toward normal because removal of CO_2 also removes carbonic acid, thereby removing hydrogen ions. Conversely, a decrease in CO_2 and hydrogen ions causes vasoconstriction, which allows these substances to accumulate in the tissues until they rise back toward the normal mean values. This is an important regulation because changes in CO_2 and hydrogen ion concentrations drastically change the degree of activity of all neurons.

LONG-TERM BLOOD FLOW REGULATION

Thus far, all the mechanisms for local blood flow regulation that we have discussed act within a few minutes after the local tissue conditions have changed. Unfortunately, though, after full function of these acute mechanisms the blood flow usually is adjusted no more than three quarters of the way to the exact requirement of the tissues. For instance, when the arterial pressure is suddenly increased from 100 mm Hg to 150 mm Hg, the blood flow increases almost instantaneously about 100 per cent. Then within one to two minutes the flow decreases back to about 15 per cent above the original control value. This illustrates the rapidity of the acute type of local regulation, but at the same time it also demonstrates that the regulation is still very incomplete because there remains a 15 per cent increase in blood flow.

However, over a period of hours, days, and weeks a *long-term type of local blood flow regulation* develops in addition to the acute regulation. This long-term regulation gives far more complete regulation than does the acute mechanism. For instance, in the above example, if the arterial pressure remains at 150 mm Hg indefinitely, within a few weeks the blood flow through the tissues will gradually reapproach almost exactly the normal value. Figure 20–5 illustrates by the dashed curve the extreme effectiveness of this long-term local blood flow regulation. Note that once the long-term regulation has had time to occur, changes in arterial pressure between 50 and 250 mm Hg have very little effect on the rate of local blood flow.

Long-term regulation also occurs when the metabolic demands of a tissue change. Thus, if a tissue becomes chronically overactive and therefore requires chronically increased quantities of nutrients, the blood supply gradually increases to match the needs of the tissue.

Long-Term Local Flow Regulation in Patients with Coarctation of the Aorta. Coarctation of the aorta means occlusion of the aorta somewhere along its course, usually in the thoracic region distal to the aortic arch. This occurs during fetal life in one of every several thousand persons, and tho person goes through life with an occluded aorta. When such an occlusion occurs, blood must flow to the lower part of the body through high resistance collateral blood vessels in the chest wall. Therefore, the arterial pressure in the upper part of the body is very high—about 1.6 times normal—whereas the arterial pressure in the lower body is almost exactly normal. Yet, despite this tremendous difference in pressure between the upper and lower body, the blood flow per unit mass of tissue is almost exactly normal both above and below the coarctation. Studies of coarctation have given us most of our quantitative information on the powerful nature of the long-term local blood flow regulatory mechanism. Indeed, this long-term mechanism seems to be many times as potent as the mechanism of acute local flow regulation

Mechanism of Long-Term Regulation—Change in Tissue Vascularity. The mechanism of long-term local blood flow regulation is almost certainly a change in the degree of vascularity of the tissues. That is, if the arterial pressure falls to 60 mm Hg and remains at this level for many weeks, the number and sizes of vessels in the tissue increase, if the pressure then rises to a very high level, the number and sizes of vessels decrease. Likewise, if the metabolism in a given tissue becomes elevated for a prolonged period of time, vascularity increases; or if the metabolism is decreased, vascularity decreases.

Thus, there is continual reconstruction of the tissue vasculature to meet the needs of the tissues. This reconstruction occurs very rapidly (within days) in extremely young animals. It also occurs rapidly in new growth of tissue, such as in scar tissue and in cancerous tissue; on the other hand, it occurs very slowly in old, well-established tissues. Therefore, the time required for long-term regulation to take place may be only a few days in the neonate or as long as months or even years in the elderly person. Furthermore, the final degree of response is much greater in younger tissues than in older, so that in the neonate the vascularity will adjust to match almost exactly the needs of the tissue for blood flow, whereas in older tissues, vascularity frequently lags far behind the needs of the tissues.

Role of Oxygen in Long-Term Regulation. A probable stimulus for increased or decreased vascularity in many if not most instances is need of the tissue for oxygen. The reason for believing this is that increased vascularity occurs in animals that live at high altitudes where the atmospheric oxygen is low. This effect has also been demonstrated dramatically in newborn premature infants who are put into an oxygen tent for therapeutic purposes. The excess oxygen causes almost immediate cessation of new vascular growth in the retina of the eye and even causes degeneration of some of the capillaries that have already formed. Then when the baby is taken out of the oxygen there is explosive overgrowth of new vessels to make up for the sudden decrease in available oxygen; indeed, there is so much overgrowth that the vessels grow into the vitreous humor and eventually cause blindness. (This condition is called *retrolental fibroplasia*.)

"Angiogenesis Factor" and Growth of New Vessels. In the past few years, a substance called *angiogenesis factor* has been isolated from several tissues that have excessive metabolic needs, including cancers and the retina of the eye. This factor causes growth of new blood vessels. Furthermore, relative lack of oxygen in relation to the level of metabolism of the tissues is one of the conditions, if not the major condition, that causes angiogenesis factor to be formed in a tissue. Therefore, much belief is developing that this angiogenesis factor is responsible for development of new blood vessels in rapidly growing, highly metabolic, or ischemic tissues

When angiogenesis factor appears in a tissue small buds of cells break away from the walls of the venules and small veins, and these cellular masses migrate toward the angiogenesis factor. Then they multiply and form tubular loops that eventually connect with the vascular tree. Thus, the new vessels are formed.

If ever enough angiogenesis factor can be isolated or synthesized, it might become an important therapeutic tool for treating ischemic tissues, such as ischemic areas in the heart after a heart attack.

Development of Collateral Circulation—A Phenomenon of Long-Term Local Blood Flow Regulation

The development of collateral circulation when blood flow to a tissue is blocked is a type of long-term local blood flow regulation Thus, if the femoral artery becomes occluded, the small vessels that bypass the femoral occlusion become greatly enlarged, and the leg, especially in younger persons, usually will develop an adequate blood supply. Within seconds after such an occlusion, blood flow to the lower leg falls to about one eighth the normal flow because the collateral vessels are still very small But acute dilatation of these vessels occurs within the first one to two minutes, so that blood flow returns to approximately one-half normal Then, over a period of a week or more the blood flow returns almost all the way to normal, indicating progressive enlargement of the collateral vessels

Thus, the opening of collateral vessels follows almost identically the same pattern as that observed in other types of acute and long-term local blood flow regulation. Furthermore, the acute opening of the collateral vessels seems to be metabolically initiated, because reduction of blood flow to a tissue by reducing arterial pressure

but without causing vascular occlusion also causes acute dilatation of the collateral vessels; that is, insufficiency of blood flow to the tissues opens the collaterals, even though the main artery is not closed.

The long-term increase in blood flow through the collaterals results from marked anatomical enlargement of the collateral vessels. The cross-sectional area of all the collaterals often becomes even greater than the original occluded artery.

Significance of Long-Term Local Regulation—The Metabolic Mass-to-Tissue Vascularity Proportionality

From these discussions it should already be apparent to the student that there is a built-in mechanism in most tissues to keep the degree of vascularity of the tissue almost exactly that required to supply the metabolic needs of the tissue. Thus, one can state as a general rule that the vascularity of most tissues of the body is directly proportional to the local metabolism If ever this proportionality constant becomes abnormal, the long-term local regulatory mechanism automatically tries to readjust the degree of vascularity over a period of weeks, months, or years. In young persons this degree of readjustment is usually very exact; in old persons it is only partial.

NERVOUS REGULATION OF THE CIRCULATION

Superimposed onto the intrinsic local tissue regulation of blood flow are two additional types of regulation: (1) *nervous* and (2) *humoral*. These regulations are not necessary for most normal functions of the circulation, but they do provide greatly increased effectiveness of control under special conditions such as during exercise or following hemorrhage.

THE AUTONOMIC NERVOUS SYSTEM

Though the autonomic nervous system will be discussed in detail in Chapter 57, it is so important to circulatory regulation that its specific anatomical and functional characteristics relating to the circulation require special attention here.

By far the most important part of the autonomic nervous system for regulation of the circulation is the *sympathetic nervous system*. However, the *parasympathetic nervous system* is also important for its regulation of heart function, as we shall see later in the chapter.

The Sympathetic Nervous System. Figure 20–6 illustrates the anatomy of sympathetic nervous control of the circulation. Sympathetic vasomotor nerve fibers leave the spinal cord through all the thoracic and the first one to two lumbar spinal nerves. These pass into the sympathetic chain and thence by two routes to the blood vessels throughout the body: (1) through specific *sympathetic nerves* that innervate mainly the internal viscera, and (2) through the *spinal nerves* that innervate mainly the peripheral areas. The precise pathways of these fibers in the spinal cord and in the sympathetic chains will be discussed in Chapter 57

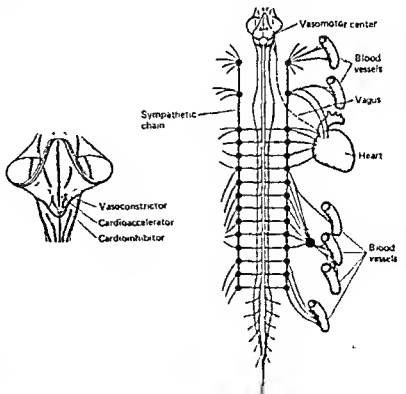


Figure 20–6. The vasomotor center and its control of the circulatory system through the sympathetic nerves and through the vagus nerves to the heart, which are part of the parasympathetic nervous system.

Sympathetic Innervation of the Vasculature. Figure 20–7 illustrates the distribution of sympathetic nerve fibers to the blood vessels and shows that all the vessels except the capillaries, precapillary sphincters, and most of the metarterioles are innervated.

The innervation of the small arteries and arterioles allows sympathetic stimulation to increase the *resistance* and thereby to change the rate of blood flow through the tissues. The innervation of large vessels, particularly of the veins, makes it possible for sympathetic stimulation to change the volume of these vessels and thereby to alter the volume of the peripheral circulatory system, which can translocate blood into the heart and thereby play a major role in the regulation of cardiovascular function, as we shall see later in this chapter and subsequent chapters.

Sympathetic Nerve Fibers to the Heart. In addition to sympathetic nerve fibers supplying the blood vessels, other sympathetic fibers go to the heart as was discussed in Chapter 13. It will be recalled that sympathetic stimulation markedly increases the activity of the heart, increasing the heart rate and enhancing its strength of pumping.

Parasympathetic Control of Heart Function, Especially Heart Rate. Though the parasympathetic nervous system is exceedingly important for many other autonomic functions of the body, it plays only a minor role in regulation of the circulation. Its only really important circulatory effect is its control of heart rate by way of parasympathetic fibers carried to the heart in the vagus nerves, shown in Figure 20–6 by the dashed nerve from the medulla directly to the heart. The parasympathetics also have a slight effect on control of cardiac contractility; however, this effect is far overshadowed by the sympathetic nervous system control of contractility.

The effects of parasympathetic stimulation on heart function were discussed in detail in Chapter 13. Principally, parasympathetic stimulation causes a marked *decrease* in heart rate and a slight decrease in contractility.

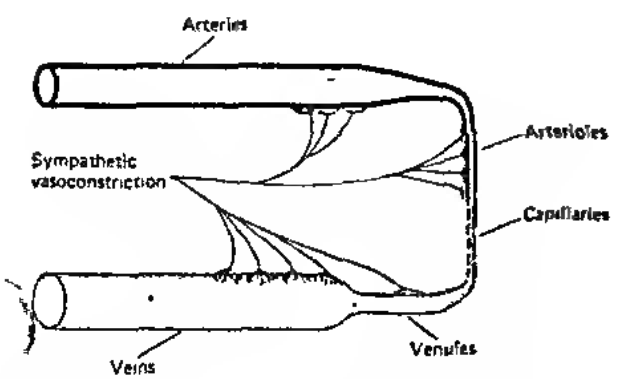


Figure 20–7. Sympathetic innervation of the systemic circulation.

The Sympathetic Vasoconstrictor System and Its Control by the Central Nervous System

The sympathetic nerves carry both vasoconstrictor and vasodilator fibers, but by far the most important of these are the *sympathetic vasoconstrictor* fibers. Sympathetic vasoconstrictor fibers are distributed to essentially all segments of the circulation. However, this distribution is greater in some tissues than in others. It is less potent in both skeletal and cardiac muscle and in the brain, whereas it is powerful in the kidneys, the gut, the spleen, and the skin.

The Vasomotor Center and Its Control of the Vasoconstrictor System—Vasomotor Tone. Located bilaterally in the reticular substance of the medulla and lower third of the pons, as illustrated in Figure 20–8, is an area called the *vasomotor center.* This center transmits impulses downward through the cord and thence through the sympathetic vasoconstrictor fibers to all or almost all the blood vessels of the body.

Though the total organization of the vasomotor center is still unclear, recent experiments have made it possible to identify certain important areas in the center, as follows:

(1) A *vasoconstrictor area*, called area "C-1," located bilaterally in the anterolateral portions of the upper medulla and lower pons. The neurons in this area secrete *norepinephrine*; their fibers are distributed throughout the cord where they excite the vasoconstrictor neurons of the sympathetic nervous system.

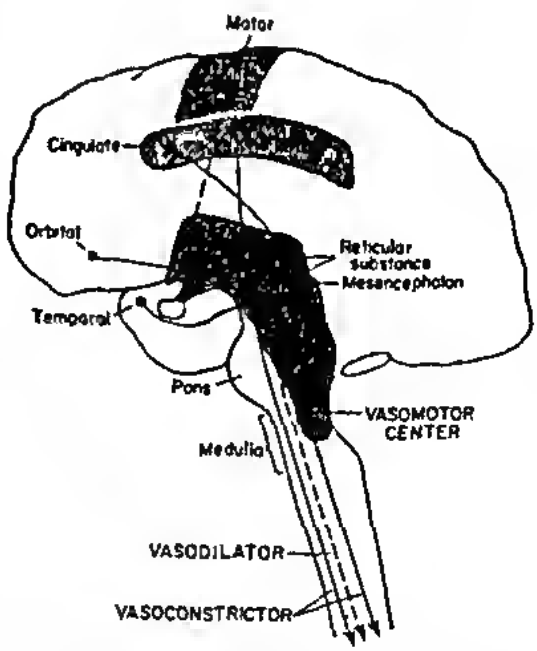


Figure 20–8. Areas of the brain that play important roles in the nervous regulation of the circulation. The dashed lines represent inhibitory pathways.

HRf MB ← VMC → Lat HRf

(2) A *vasodilator area*, called area "A-1," located bilaterally in the anterolateral portions of the lower half of the medulla. The neurons in this area also secrete *norepinephrine*, but their fibers project upward to the vasoconstrictor area (A-1) and inhibit the vasoconstrictor activity of that area, thus causing vasodilation.

(3) A *sensory area*, located bilaterally in the *tractus solitarius* in the posterolateral portions of the medulla and lower pons. The neurons of this area receive nerve signals mainly from the vagus and glossopharyngeal nerves, and the output signals from the sensory area then help control the activities of both the vasoconstrictor and vasodilator areas, thus providing sensory control of many of the vasomotor center's activities.

Under normal conditions, the vasoconstrictor area of the vasomotor center transmits signals continuously to the sympathetic vasoconstrictor nerve fibers, causing continuous slow firing of these fibers at a rate of about one half to two impulses per second. This continual firing is called *sympathetic vasoconstrictor tone*. These impulses maintain a partial state of contraction in the blood vessels, a state called *vasomotor tone*.

Figure 20–9 demonstrates the significance of vasoconstrictor tone. In the experiment of this figure, total spinal anesthesia was administered to an animal, which completely blocked all transmission of nerve impulses from the central nervous system to the periphery. As a result, the arterial pressure fell from 100 to 50 mm Hg, illustrating the effect of loss of vasoconstrictor tone throughout the body. A few minutes later a small amount of the hormone norepinephrine was injected intravenously—norepinephrine is the substance *secreted at the endings of sympathetic nerve fibers* throughout the body. As this hormone was transported in the blood to all the blood vessels, the vessels once again became constricted, and the arterial pressure rose to a level even greater than normal for a minute or two until the norepinephrine was destroyed.

Control of Heart Activity by the Vasomotor Center. At the same time that the vasomotor center is controlling the degree of vascular constriction, it also controls heart activity. The lateral portions of the vasomotor center transmit excitatory impulses through the sympathetic nerve fibers to the heart to increase heart rate and contractility, while the medial portion of the vasomotor center, which lies in immediate apposition to the *dorsal motor nucleus of the vagus nerve*, transmits impulses through the vagus nerve to the heart to decrease heart rate. Therefore, the vasomotor center can either increase or decrease heart activity, this ordinarily increasing at the same time that vasoconstriction occurs throughout the body and ordinarily decreasing at the same time that vasoconstriction is inhibited. However, these interrelationships are not invariable, because some signals that pass down the vagus nerves to the heart can bypass the vasomotor center.

Control of the Vasomotor Center by Higher Nervous Centers. Large numbers of areas throughout the *reticular substance* of the *pons*, *mesencephalon*, and *diencephalon* can either excite or inhibit the vasomotor center. This reticular substance is illustrated in Figure 20–8 by the diffuse shaded area. In general, the more lateral and superior portions of the reticular substance cause excitation, while the more medial and inferior portions cause inhibition.

The *hypothalamus* plays a special role in the control of the vasoconstrictor system, for it can *exert powerful excitatory or inhibitory effects on* the vasomotor center. The *posterolateral portions* of the hypothalamus cause mainly excitation,

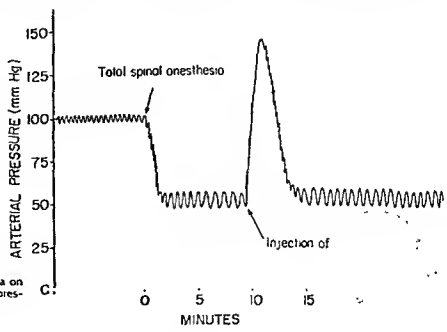


Figure 20–9. Effect of total spinal anesthesia on the arterial pressure, showing a marked fall in pressure resulting from loss of vasomotor tone.

while the anterior part can cause mild excitation or inhibition, depending on the precise part of the anterior hypothalamus stimulated.

Many different parts of the *cerebral cortex* can also excite or inhibit the vasomotor center. Stimulation of the *motor cortex*, for instance, excites the vasomotor center because of impulses transmitted downward into the hypothalamus and thence to the vasomotor center. Also, stimulation of the *anterior temporal lobe*, the *orbital areas of the frontal cortex*, the *anterior part of the cingulate gyrus*, the *amygdala*, the *septum*, and the *hippocampus* can all either excite or inhibit the vasomotor center, depending on the precise portion of these areas that is stimulated and upon the intensity of the stimulus

Thus, widespread areas of the brain can have profound effects on the vasomotor center and in turn on the sympathetic vasoconstrictor system of the body, either further enhancing the degree of vasoconstriction or causing vasodilatation by inhibiting the vasoconstrictor tone.

Norepinephrine—The Sympathetic Vasoconstrictor Transmitter Substance. The substance secreted at the endings of the vasoconstictor nerves is norepinephrine. Norepinephrine acts directly on the so-called "alpha" receptors of the vascular smooth muscle to cause vasoconstriction, as will be discussed in Chapter 57.

The Adrenal Medullae and Their Relationship to the Sympathetic Vasoconstrictor System. Sympathetic impulses are transmitted to the adrenal medullae at the same time that they are transmitted to all the blood vessels. These impulses cause the medullae to secrete both epinephrine and norepinephrine into the circulating blood, as will be described in Chapter 57. These two hormones are carried in the bloodstream to all parts of the body, where they act directly on the blood vessels usually to cause vasoconstriction, but sometimes the epinephrine causes vasodilatation because it has a potent "beta" receptor stimulatory effect, which often dilates vessels, as will also be discussed in Chapter 57

The Sympathetic Vasodilator System and Its Control by the Central Nervous System

The sympathetic nerves to skeletal muscles carry sympathetic vasodilator fibers as well as constrictor fibers In lower animals, such as the cat, these fibers release *acetylcholine*, not norepinephrine, at their endings, though in primates the vasodilator effect might be caused by epinephrine exciting the beta receptors in the muscle vasculature.

The pathways for central nervous system control of the vasodilator system are illustrated by the dashed [illegible] in Figure 20–8 The principal area of the brain controlling this system is the *anterior hypothalamus*, which transmits impulses to a relay station in the *subcollicular region* of the mesencephalon From here impulses are transmitted down the cord to *sympathetic neurons* in the lateral horns of the cord. These then transmit vasodilator signals to the blood vessels of the muscles

Note also in Figure 20–8 that stimulation of the *motor cortex excites* the *sympathetic vasodilator system* by transmitting impulses downward into the hypothalamus

Importance of the Sympathetic Vasodilator System. It is doubtful that the sympathetic vasodilator system plays a very important role in the control of the circulation, because complete block of the sympathetic nerves to the muscles hardly affects the ability of the muscles to control their own blood flow in response to their needs Yet, it is possible, if not probable, that at the onset of exercise the sympathetic vasodilator system causes initial vasodilation in the skeletal muscles to allow an *anticipatory increase in blood flow* even before the muscles require increased nutrients.

"PATTERNS" OF CIRCULATORY RESPONSES ELICITED BY DIFFERENT CENTRAL NERVOUS SYSTEM CENTERS

Stimulation of the Vasomotor Center—The Mass Action Effect. Stimulation of the lateral portions of the vasomotor center causes widespread activation of the vasoconstrictor fibers throughout the body, while stimulation of the medial portions of the vasomotor center causes widespread *inhibition of vasoconstriction* In many conditions the entire vasomotor center acts as a unit, stimulating at the same time all vasoconstrictors throughout the body and also the heart, as well as stimulating the adrenal medullae to secrete epinephrine and norepinephrine that circulate in the blood to excite the circulation still further The results of this "mass action" are threefold First, the peripheral resistance increases in most parts of the circulation, thereby elevating the arterial pressure Second, the capacity vessels, particularly the veins, are excited, which greatly decreases their capacity; this forces increased quantities of blood into the heart, thereby increasing the cardiac output Third, the heart is strongly stimulated so that it can handle the increased cardiac output.

Thus, the overall effect of this "mass action" is to prepare the circulation for increased delivery of blood flow to the body

Stimulation of the Hypothalamus—The "Alarm" Pattern. Under normal conditions, the hypothalamus probably does not transmit large numbers of impulses into the sympathetic vascular control system However, on occasion the hypothalamus becomes strongly stimulated and can then activate either the vasoconstrictor or the vasodilator systems or both of them together

Diffuse stimulation of the hypothalamus (or of closely allied areas such as the septum) activates the vasodilator system to the muscles, thereby increasing the blood flow through the muscles; at the same time it causes intense vasoconstriction throughout the remainder of the body and an intense increase in heart activity. The arterial pressure rises, the cardiac output increases, the heart rate increases, and the circulation is ready to supply nutrients to the muscles if there be need Also, impulses are transmitted simultaneously throughout the central nervous system to cause a state of generalized excitement and attentiveness, these often increasing to such a pitch that the overall pattern of the reaction is that of *alarm*, or sometimes also called the *defense pattern*.

Thus, the "alarm" pattern contains all the ingredients of the "mass action" pattern as well as the elements of muscle vasodilation and extreme psychic excitement. This pattern seems to have the purposeful effect of preparing the animal or person to perform on a second's notice whatever activity is required.

The "Motor" Pattern of Circulatory Stimulation. When the motor cortex transmits signals to the skeletal muscle to cause motor activities, it also sends signals to the circulatory system to cause the following effects: First, it excites the alarm system just described, and this generally activates the heart and increases arterial pressure, making the circulatory system ready to supply increased blood flow to the muscles. This effect also causes vasoconstriction in most nonmuscular parts of the body, such as the kidneys, the skin, and the gut, thereby forcing a larger share of the blood flow through the active muscles. Second, impulses pass directly from the motor cortex to the sympathetic neurons of the spinal cord. These enhance the vasoconstriction in the nonmuscular parts of the body, thereby raising the arterial pressure still more and helping to increase muscle blood flow. They possibly also inhibit the vasoconstrictor nerves to the excited muscles.

The motor pattern of circulatory response can also be elicited by stimulation of several areas of the reticular substance in the brain stem. Stimulation of one of these areas, the *fields of Forel*, elicits a response almost precisely the same as that caused by direct activation of the motor cortex. Therefore, it is possible, if not likely, that this area of the reticular substance plays a major role in control of the circulatory system in exercise

Emotional Fainting—Vasovagal Syncope. A particularly interesting circulatory reaction occurs in persons who faint because of intense emotional experiences. In this condition, the muscle vasodilator system becomes powerfully activated so that blood flow through the muscles increases severalfold. Intense vagal stimulation of the heart also occurs, causing the heart rate to slow markedly. This overall effect is called *vasovagal syncope* The arterial pressure falls instantly, which in turn reduces the blood flow to the brain and causes the person to lose consciousness It is probable, therefore, that emotional fainting results from powerful stimulation of the anterior hypothalamic vasodilator center.

REFLEX REGULATION OF THE CIRCULATION

In addition to the many ways in which signals transmitted from the brain to the blood vessels and heart can alter circulatory function, a system of specific circulatory reflexes also helps to regulate the circulation. Most of these reflexes are concerned with the regulation of arterial pressure and therefore will be discussed in the following chapter in relation to arterial pressure regulation. However, let us summarize here a few of the more important reflex regulations of the circulation.

Arterial Pressure Reflexes. The most important of the arterial pressure reflexes is the *baroreceptor reflex*. An increase in arterial pressure stretches the walls of the major arteries in the chest and neck, and this, in turn, excites stretch receptors, the *baroreceptors*. Signals are transmitted to the vasomotor center of the brain stem, and reflex signals are transmitted back to the heart and blood vessels to slow the heart and to dilate the vessels, thereby reducing the arterial pressure toward normal. Thus, the baroreceptor reflex helps to stabilize the arterial pressure.

Reflexes for Control of Blood Volume. When the blood volume increases, the volume of blood in the central veins and in the right and left atria of the heart usually also increases This stretches the walls of the atria and large veins, again exciting stretch receptors. These stretch receptors also transmit signals into the vasomotor center, and reflex signals are then transmitted to the kidneys to increase urinary output of fluid Also, signals are transmitted to the hypothalamus to decrease the rate of secretion of antidiuretic hormone, thereby also increasing the output of fluid by the kidneys. Thus, this reflex mechanism helps to control the total amount of fluid in the body, and in this way also helps to control blood volume

Nervous Control of Body Temperature. When the body temperature rises too high, special neurons in the anterior hypothalamus become excited; these in turn send signals through the sympathetic nervous system to dilate the skin blood vessels, thereby allowing transfer of major amounts of internal body heat to the skin This heat is then eliminated from the skin to the surroundings. As a result, the body temperature falls back toward normal. Other nervous signals at the same time control sweating and other aspects of body temperature control, all of which will be discussed in Chapter 72

HUMORAL REGULATION OF THE CIRCULATION

Humoral regulation of the circulation means regulation by substances in the body fluids, such as by hormones, ions, or so forth. Some of these substances are formed by special glands and then transported in the blood throughout the entire body. Others are formed in local tissue areas and cause only local circulatory effects. Among the *most important of the humoral factors that affect* circulatory function are the following:

Vasoconstrictor Agents

Epinephrine and Norepinephrine. Earlier in the chapter it was pointed out that the adrenal medullae secrete both epinephrine and norepinephrine, usually more of the former than the latter. When the sympathetic nervous system is stimulated to cause direct effects on the blood vessels throughout the body, it also causes the adrenal medullae to secrete these two hormones, which then circulate everywhere in the body fluids and act on all vasculature. Norepinephrine has vasoconstrictor effects in almost all vascular beds of the body, and epinephrine has similar effects in most, but not all, beds. For instance, epinephrine often causes mild vasodilatation in both skeletal and cardiac muscle. These two hormones will be discussed in more detail in Chapter 57 in the discussion of the autonomic nervous system.

Angiotensin. Angiotensin is one of the most powerful vasoconstrictor substances known. As little as *one ten millionth* of a gram can increase the arterial pressure of a human being as much as 10

to 20 mm Hg under some conditions. Since this substance is very important in relation to arterial pressure regulation, it will be discussed in detail in the following two chapters.

Briefly, a decrease in arterial pressure will cause the kidneys to secrete the substance *renin*. The renin in turn acts on one of the plasma proteins, *renin substrate*, to split away the vasoactive peptide *angiotensin* The angiotensin in turn has a number of important effects on the circulation related to arterial pressure control, but most importantly to constrict the blood vessels everywhere in the body, as we shall discuss in the following chapters.

Vasopressin. Vasopressin, also called *antidiuretic hormone*, is formed in the hypothalamus (see Chapter 75) but is transported down the center of nerve axons to the posterior pituitary gland, where it is eventually secreted into the blood. It is even more powerful than angiotensin as a vasoconstrictor, thus making it perhaps the body's most potent constrictor substance. Therefore, it is clear that vasopressin could have very intense effects on circulatory function. Yet, normally, only very minuta amounta of vasopressin are secreted, so that most physiologists have thought that vasopressin playa little role in vascular control. On the other hand, recent experiments have shown that the concentration of circulating vasopressin during severe hemorrhage can rise enough to increase the arterial pressure as much as 60 mm Hg, and in many instances this can by itself bring the arterial pressure almost back up to normal.

Also, vasopressin has an *all-important* function in controlling water reabsorption in the renal tubules, which will be discussed in Chapter 36, and therefore to help control body fluid volume That is why this hormone is also called antidiuretic hormone.

Vasodilator Agents

Bradykinin. Several substances called *kinins* that can cause vasodilatation have been isolated from blood and tissue fluids One of these substances is *bradykinin*

The kinins are small polypeptides that are split away from alpha$_2$-globulins in the plasma or tissue fluids. Different typea of proteolytic enzymes can split the kinins from the globulin. An enzyme of particular importance is *kallikrein*, which is present in the blood and tissue fluids in an inactive form. Kallikrein can be activated in several different ways, such as by maceration of the blood, dilution of the blood, contact of the blood with glass, and other similar chemical and physical effects on the blood. As kallikrein becomea activated, it acts immediately on the alpha$_2$-globulin to release a [illegible] called *kallidin* that is then converted by tissue [illegible]zymes into *bradykinin* Once formed, the bradykinin [illegible] for only a few minutes, because it is digested by the enzyme *carboxypeptidase* or by *converting enzyme* that also playa an essential role in activating angiotensin, aa discussed in Chapter 21. The activated kallikrein enzyme is destroyed by a kallikrein inhibitor also present in the body fluids

Bradykinin causea very powerful *arteriolar dilatation* and also *increased capillary permeability*. For instance, injection of 1 *microgram* of bradykinin into the brachial artery of a man increases the blood flow through the arm as much as sixfold, and even smaller amounts injected locally into tissues can cause marked edema because of the increase in capillary pore size.

Unfortunately, though, we know little about the function of the kinins in the control of the circulation. Nevertheless, these extremely powerful effects, coupled with the fact that kinina can develop anywhere in the circulatory aystem or tissues with ease, indicate that they must play important roles in circulatory regulation. There is reason to believe that kinins play special roles in regulating blood flow and capillary leakage of fluids in inflamed tissues It is also believed that bradykinin plays a role in regulating blood flow in the skin and also in the salivary and gastrointestinal glands.

Serotonin. Serotonin (5-hydroxytryptamine) is present in large concentrations in the chromaffin tissue of the intestine and other abdominal structures. Also, it is present in high concentration in the platelets Serotonin can have either a vasodilator or vasoconstrictor effect, depending on the condition or the area of the circulation. And, even though these effects can sometimes be powerful, the functions of serotonin in regulation of the circulation are almost entirely unknown. Occasionally, tumors composed of chromaffin tissue develop, called *carcinoid tumors*. These secrete tremendous quantities of serotonin and cause mottled areas of vasodilatation in the skin; but the very fact that thesa tremendous quantities of serotonin do not drastically disturb the circulation makes it doubtful that serotonin plays a widespread general role in regulation of circulatory function

Histamine. Histamine is released in essentially every tissue of the body whenever it becomes damaged. Most of the histamine is probably derived from mast cells in the damaged tissues and from basophils in the blood.

Histamine has a powerful vasodilator effect on the arterioles, and like bradykinin, also has the ability to greatly increase capillary porosity, allowing leakage of both fluid and plasma protein into the tissues. Though the role of histamine in normal regulation of the circulation is unknown, in many pathological conditions the intense arteriolar dilatation and increased capillary porosity caused by histamine cause tremendous quantities of fluid to leak out of the circulation into the tissues, inducing edema The effects of histamine were discussed in Chapter 7 in relation to allergic reactions

Prostaglandins. Almost every tissue of the body contains small to moderate amounts of several chemically related substances called prostaglandins. These substances are released into the local tissue fluids and into the circulating blood under both physiological and pathological conditions Though some of the prostaglandins cause vasoconstriction, most of the more important ones seem to be mainly vasodilator agents Thus far, no specific pattern of function of the prostaglandins in circulatory control has been found However, their widespread prevalence in the tissues and their myriad effects on the circulation make them ideal candidates for special roles in circulatory control, especially for control in local vascular areas For this reason these substances are presently under intensive research investigation,

though unequivocal conclusions about their roles in circulatory regulation are still evasive.

Effects of Different Ions and Other Chemical Factors on Vascular Control

Many different ions and other chemical factors can either dilate or constrict local blood vessels. Though the roles of these substances in the overall *regulation* of the circulation generally are not known, their specific effects can be listed as follows:

An increase in *calcium ion* concentration causes vasoconstriction. This results from the general effect of calcium to stimulate smooth muscle contraction, as discussed in Chapter 12.

An increase in *potassium ion* concentration causes vasodilatation. This results from the ability of potassium ions to inhibit smooth muscle contraction.

An increase in *magnesium ion concentration* causes powerful vasodilatation, for magnesium ions inhibit smooth muscle generally.

Increased *sodium ion* concentration causes arteriolar dilatation. This results mainly from an increase in osmolality of the fluids rather than from a specific effect of sodium ion itself. *Increased osmolality* of the blood caused by increased quantities of *glucose* or other non-vasoactive substances also causes arteriolar dilatation. Decreased osmolality causes arteriolar constriction.

The only anions to have significant effects on blood vessels are *acetate* and *citrate*, both of which cause mild degrees of vasodilatation.

An *increase in hydrogen* ion concentration (decrease in pH) causes dilatation of the arterioles. A slight *decrease in hydrogen ion* concentration causes arteriolar constriction, but an intense decrease causes dilatation, which is the same effect as that which occurs with increased hydrogen ion concentration

An increase in carbon dioxide concentration causes moderate vasodilatation in most tissues and marked vasodilatation in the brain. However, carbon dioxide, acting on the vasomotor center, has an extremely powerful indirect vasoconstrictor effect that is transmitted through the sympathetic vasoconstrictor system.

REFERENCES

Abboud, F. M (ed) Disturbances in Neurogenic Control of the Circulation. Baltimore, The Williams & Wilkins Co, 1981

American Physiological Society Peripheral Circulation and Organ Blood Flow Washington, D C, American Physiological Society, 1983

American Physiological Society Microcirculation. Washington, D C, American Physiological Society, 1984

Antonaccio, M J. (ed) Cardiovascular Pharmacology New York, Raven Press, 1984

Bevan, J. A., *et al* Adrenergic regulation of vascular smooth muscle *In* Bohr, D F, *et al.* (eds) Handbook of Physiology Sec. 2 Vol II Baltimore, Williams & Wilkins, 1980, p 515

Bevan, J. A, *et al.* (eds). Vascular Neuroeffector Mechanisms. New York, Raven Press, 1980

Bhoola, K. D, and Erdös, E. G (eds) Bradykinin, Kallidin, and Kallikrein. New York, Springer-Verlag, 1979

Borgstrom, P., *et al.*. An evaluation of the metabolic interaction with myogenic vascular reactivity during blood flow autoregulation *Acta Physiol Scand*, *122* 275, 1984

Buhler, F. R (ed) Alpha-Adrenoceptors in Cardiovascular Regulation New York, Raven Press, 1982.

Burnstock, G.: Cholinergic and purinergic regulation of blood vessels. *In* Bohr, D F, *et al.* (eds). Handbook of Physiology Sec. 2, Vol. II Baltimore, Williams & Wilkins, 1980, p 567.

Cowley, A. W, Jr, and Cuyton, A. C.: Quantification of intermediate steps in the renin angiotensin vasoconstrictor feedback loop in the dog. *Circ. Res.*, *30* 557, 1972

Dampney, R. A, *et al* Identification of cardiovascular cell groups in the brain stem *Chn. Exp Hypertens*, *6* 205, 1984

Duling, B: Oxygen, metabolism, and microcirculatory regulation *In* Kaley, G, and Altura, B M (eds) Microcirculation, Vol II Baltimore, *University Park Press*, 1977

Duling, B R, and Kuitzman, B Local control of microvascular function role in tissue oxygen supply *Annu Rev Physiol*, *42* 373, 1980

Folkman, J The vascularization of tumors *Sci Am*, *234*(5) 58, 1978

Folkow, B. Physiological aspects of primary hypertension *Physiol Rev*, *62* 347, 1982.

Granger, H. J, and Guyton, A. C Autoregulation of the total systemic circulation following destruction of the central nervous system in the dog. *Circ Res*, *25* 379, 1969

Guyton, A C Integrative hemodynamics *In* Sodeman, W A, Jr., and Sodeman, T M (eds) Pathologic Physiology Mechanisms of Disease, 6th Ed Philadelphia, W B Saunders Co 1979, p 168

Guyton, A C *Arterial* Pressure and Hypertension Philadelphia, W B Saunders Co, 1980.

Guyton, A C, *et al*, Circulation Overall regulation. *Annu Rev Physiol*, *34* 13, 1972.

Guyton, A. C., *et al* Cardiac Output and Its Regulation. Philadelphia, W B. Saunders Co, 1973

Haddy, F J. Local effects of sodium, calcium, and magnesium upon small and large blood vessels of the dog forelimb *Circ Res*, *8* 57, 1960

Haddy, F J. The role of a humoral Na+, K+ – ATPase inhibitor in *regulating precapillary vessel tone J Cardiovasc Pharmacol*, 6(Suppl 2)S439, 1984

Hegyeli, R J (ed) Prostaglandins and Cardiovascular Disease New York, Raven Press, 1981

Hudlicka, O Development of microcirculation Capillary growth and *adaptation In* Renkin, E M, *and Michel*, C C *(eds)* Handbook of Physiology Sec 2, Vol IV Bethesda, American Physiological Society, 1984, p 165

Johansson, B Vascular smooth muscle reactivity *Annu Rev Physiol*, *43* 359, 1981

Johnson, P C The myogenic response *In* Bohr, D F, *et al (eds)* Handbook of Physiology Sec 2 Vol II Baltimore, Williams & Wilkins, 1980, p 409

Keatinge, W R., and Harman, M C Local Mechanisms Controlling Blood Vessels New York, Academic Press, 1979

Kreye, V A. Direct vasodilators with unknown *modes of action*. The nitro-compounds and hydralazine *J Cardiovasc Pharmacol*, 6(Suppl 4)S646, 1984

Mitchell, J H, *et al* The exercise pressor reflex. Its cardiovascular effects, afferent mechanisms, and control pathways. *Annu. Rev Physiol*, *45* 229 1983

Olsson, R A Local factors regulating cardiac and skeletal muscle blood flow *Annu Rev Physiol*, *43* 385, 1981

Randall, W C (ed) Nervous Control of Cardiovascular Function. New York, Oxford University Press, 1984

Regoli, D Neurohumoral regulation *of precapillary* vessels. The kallikrein-kinin system *J Cardiovasc Pharmacol*, 6(Suppl 2)S401, 1984

Renkin, E M Control of microcirculation and blood tissue exchange *In* Renkin, E M, and Michel, C C (eds.) Handbook of Physiology Sec 2, Vol IV Bethesda, American Physiological Society, 1984, p 627

Renkin, E M Nutritiva and shunt flow *In* Kaley, G, and Altura, B M (eds) Microcirculation Vol II Baltimore, University Park Press, 1977

Rodiewicz, C M Arteries and Arterial Blood Flow New York, *Springer-Verlag*, 1983

Rosell, S Neuronal control of microvessels *Annu Rev Physiol*, *42* 359, 1980.

Rosenthal, S L, and Cuyton, A. C: Hemodynamics of collateral vasodilatation following femoral artery occlusion in anesthetized dogs. *Circ Res*, *23* 239, 1968

Rowell, L. B.. Reflex control of regional circulation in humans. *J. Autonom Nerv Syst*, *11* 101, 1984

Silverman, W A. The lesion of retrolental fibroplasia. *Sci Am*, 236(6) 100, 1977

Sparks, H V, Jr. Effect of local metabolic factors on vascular smooth muscle. *In* Bohr, D F, *et al.* (eds) Handbook of Physiology. Sec. 2, Vol. II. Baltimore, Williams & Wilkins, 1980, p. 475

Vallee, E L., *et al.*. Tumor-derived angiogenesis factors from rat Walker 256 carcinoma An experimental investigation and review. *Experienta*, *41*.1, 1985

Vanhoutte, P. M: Physical factors of regulation. *In* Bohr, D F, *et al.* (eds) Handbook of Physiology Sec. 2, Vol II Baltimore, Williams & Wilkins, 1980, p 443

Vanhoutte, P M, *et al.*: Local modulation of the adrenergic neuroeffector interaction in the blood vessel wall. *Physiol. Rev.*, *61*.131, 1981

Wyse, G. D, *et al.*: Responses of human peripheral arteries to agonists Similarities and differences. *Fed. Proc*, *44*.331, 1985

21

Arterial Pressure Regulation: I. Rapid Pressure Control by Nervous Reflexes and Other Mechanisms

In the previous chapter we pointed out that each tissue can control its own blood flow by simply dilating or constricting its local arterioles For this mechanism to work, it is necessary that the arterial pressure remain constant or nearly constant because with a variable arterial pressure one would never know whether dilating the blood vessels would necessarily increase the local blood flow Fortunately, the circulation has an intricate system for regulation of the arterial pressure It maintains the normal mean arterial pressure in the young adult within rather narrow limits between 95 mm Hg and 100 mm Hg Some of the pressure regulatory mechanisms (mainly nervous and hormonal mechanisms) act very rapidly, and some (mainly mechanisms related to kidney function and blood volume regulation) act very slowly. In the present chapter we will discuss the rapid nervous and hormonal pressure control mechanisms In the following chapter we will discuss both the long-term regulation of arterial pressure, based primarily on renal and body fluid mechanisms, and the clinical problem of hypertension or "high blood pressure," which is caused by abnormalities of the long-term pressure regulatory mechanisms.

NORMAL ARTERIAL PRESSURES

Arterial Pressures at Different Ages. Figure 21-1 ustrates the typical diastolic, systolic, and mean ar- pressures from birth to 80 years of age From this gure it can be seen that the systolic pressure of a normal young adult averages about 120 mm Hg and the diastolic pressure about 80 mm Hg—that is, arteria pressure is said to be 120 80 The shaded areas on eithe side of the curves depict the normal ranges of systoli and diastolic pressures, showing considerable variatio from person to person

The increase in arterial pressure at older ages i usually associated with developing arteriosclerosis. I this disease the systolic pressure especially increases in approximately one-tenth of all old people it eventually rises above 200 mm Hg

THE MEAN ARTERIAL PRESSURE

The mean arterial pressure is the average pressure throughout each cycle of the heartbeat Offhand one might expect that it would be equal to the average o systolic and diastolic pressures but this is not true; the arterial pressure usually remains nearer to diastolic level than to systolic level during a greater portion of the pulse cycle, which can be seen in all the pictures of pressure pulses shown in Chapter 19 Therefore, the mean arterial pressure is usually slightly less than the average of systolic and diastolic pressures, as is evident in Figure 21-1.

Mean Arterial Pressure in the Human Being. The mean arterial pressure of the normal young adult averages about 96 mm Hg, which is slightly less than the average of the systolic and diastolic pressures, 120 and 80 mm Hg, respectively. However, for purposes of discussion, the mean arterial pressure is usually considered to be 100 mm Hg because this value is easy to remember

The mean arterial pressure, like the systolic and diastolic pressures, is lowest immediately after birth, measuring about 70 mm Hg at birth and reaching an average of about 110 mm Hg in the normal old person or as high as 130 mm Hg in the person with arteriosclerosis. From adolescence to middle age the mean arterial

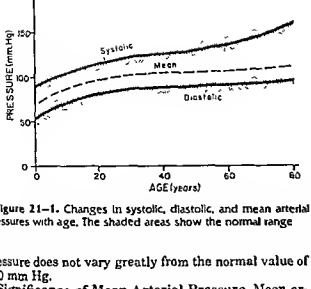


Figure 21–1. Changes in systolic, diastolic, and mean arterial pressures with age. The shaded areas show the normal range

pressure does not vary greatly from the normal value of 100 mm Hg.

Significance of Mean Arterial Pressure. Mean arterial pressure is the average pressure tending to push blood through the systemic circulatory system Therefore, *from the point of view of tissue blood flow, it is generally the mean arterial pressure that is important.*

CLINICAL METHODS FOR MEASURING SYSTOLIC AND DIASTOLIC PRESSURES

Obviously, it is impossible to use the various pressure recorders that require needle insertion into an artery, as described in Chapter 18, for making routine pressure measurements in human patients, although they are used on occasion when special studies are necessary Instead, the clinician determines systolic and diastolic pressures by indirect means, most usually by the auscultatory method.

The Auscultatory Method. Figure 21–2 illustrates the auscultatory method for determining systolic and diastolic arterial pressures A stethoscope is placed over the antecubital artery while a blood pressure cuff is inflated around the upper arm. As long as the cuff

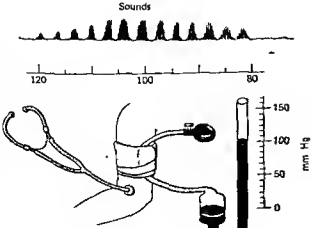


Figure 21–2. The auscultatory method of measuring systolic and diastolic pressures.

presses against the arm with so little pressure that the artery remains distended with blood, no sounds whatsoever are heard by the stethoscope despite the fact that the blood within the artery is pulsating. But when the cuff pressure is great enough to close the artery during part of the arterial pressure cycle, a sound is heard in the stethoscope with each pulsation. These sounds are called *Korotkoff sounds.*

The exact cause of Korotkoff sounds is still debated, but they are believed to be caused by blood jetting through the partly occluded vessel The jet causes turbulence in the open vessel beyond the cuff, and this sets up the vibrations heard through the stethoscope

In determining blood pressure by the auscultatory method, the pressure in the cuff is first elevated well above arterial systolic pressure As long as this pressure is higher than systolic pressure, the brachial artery *remains collapsed and no blood whatsoever flows into* the lower artery during any part of the pressure cycle Therefore, no Korotkoff sounds are heard in the lower artery *But then the cuff pressure is gradually reduced* Just as soon as the pressure in the cuff falls below systolic pressure, blood slips through the artery beneath the cuff during the peak of systolic pressure, and one begins to hear *tapping* sounds in the antecubital artery in synchrony with the heart beat As soon as these sounds are heard, this pressure level indicated by the manometer connected to the cuff is approximately equal *to the systolic pressure.*

As the pressure in the cuff is lowered still more, the Korotkoff sounds change in quality, having less of the *tapping quality but more of a rhythmic, harsher quality.* Then, finally, when the pressure in the cuff falls to equal diastolic pressure, the artery no longer closes during diastole, which means that the basic factor causing the sounds (the jetting of blood through a squeezed artery) is no longer present Therefore, the sounds suddenly change to a muffled quality, and they usually disappear entirely after another 5 to 10 mm drop in cuff pressure One notes the manometer pressure when the Korotkoff sounds change to the muffled quality, and this pressure is approximately equal to the diastolic pressure

The auscultatory method for determining systolic and diastolic pressures is not entirely accurate, but it usually gives values within 10 per cent of those determined by *direct measurement from the arteries*

Radial Pulse and Oscillometric Methods for Estimating Arterial Pressure. Arterial blood pressure can also be estimated *by feeling the radial pulse* or by recording the pulsation in the lower arm with an *oscillometer* while a cuff is inflated over the upper arm. An oscillometer consists of a recording apparatus that can register pulsations in a lightly inflated blood pressure cuff around the forearm. A blood pressure recorder of the aneroid type makes an excellent oscillometer for estimating arterial pressure.

In applying these methods, the pressure in the cuff on the upper arm is raised well above the systolic level and then is progressively reduced. No pulsation will be felt in the radial artery or registered by the oscillometer until the cuff pressure falls below the systolic pressure level. But, just as soon as this point is reached, a distinct radial pulse can be felt, or a distinct oscillation will be recorded by the oscillometer. Further decrease in the upper arm cuff pressure causes, at first, a progressive *increase in intensity of radial or oscillometer pulsation.* Then at approximately the diastolic level of pressure the intensity of the pulsations decreases slightly.

ever, this change at the diastolic pressure level is so indistinct that estimation of diastolic pressure is likely to be in severe error

The oscillometric method is valuable for measuring pressures when distinct sounds cannot be heard from the forearm arteries This often occurs in (1) very small children or (2) adults when the arteries are in a state of spasm, such as when a patient is in *shock* These instances are often among the most important in which it is essential to have at least some estimation of arterial pressure, thus illustrating the value of knowing this accessory method for measuring systolic arterial pressure in addition to the auscultatory method.

RELATIONSHIP OF ARTERIAL PRESSURE TO CARDIAC OUTPUT AND TOTAL PERIPHERAL RESISTANCE

Before discussing the overall regulation of arterial pressure, it is good to remember the basic relationship between arterial pressure, cardiac output, and total peripheral resistance, which was discussed in detail in Chapter 18, as follows:

$$\text{Arterial pressure} = \text{Cardiac output} \times \text{total peripheral resistance}$$

It is obvious from this formula that any condition that increases either the cardiac output or total peripheral resistance (if the other factor does not change) will cause an increase in mean arterial pressure. Both these factors are often manipulated in the control of arterial pressure, as we shall see in the remainder of this chapter and in the following chapter

THE OVERALL SYSTEM FOR ARTERIAL PRESSURE REGULATION

Arterial pressure is not regulated by a single pressure controlling system but instead by several interrelated systems that perform specific functions. When a person bleeds severely so that the pressure falls suddenly, two problems immediately confront the pressure control system The first ia to return the arterial pressure immediately to a high enough level that the person can live through the acute hemorrhagic episode. The second is to return the blood volume eventually to its normal level so that the circulatory system can re-establish full normality, including return of the arterial pressure all the way back to its normal value, not merely back to a pressure level required for survival These two problems characterize two major y. of arterial pressure control systems in the body: (1) a system of rapidly acting pressure control mechanisms that are concerned with immediate survival, and (2) a system for long-term control of the basic arterial pressure level.

To illustrate the multiplicity of pressure cont mechanisms, let us look for a moment at Fi 21–3. This figure depicts the (expressed as feedback gain) of eight arterial pressure control systems, showing a times required for them to begin to respond a their durations of response, following an acu abnormal change in the arterial pressure. The control mechanisms can be divided into those th react very rapidly and those that provide lo term arterial pressure regulation.

Rapidly-Acting Pressure Control Mech nisms. Note in Figure 21–3 that three differe pressure control mechanisms begin to react with seconds. All are nervous pressure control mech nisms: (a) the baroreceptor feedback mechanis (b) the central nervous system ischemic mech nism, and (c) the chemoreceptor mechanism. Th the first line of defense against abnormal pressu is the nervous mechanisms for control of arter pressure

Within minutes several other pressure cont mechanisms also come into play. Three of the also shown in Figure 21–3, are (1) the ren angiotensin-vasoconstrictor mechanism, (2) str relaxation changes in the vasculature, and (3) sh of fluid through the capillaries from the tiss into or out of the circulation to readjust the blo volume as needed. These three mechanisms come fully active within 30 minutes to seve hours, in contrast to the nervous mechanisms th usually become fully active within a minute or

Long-Term Mechanism for Arterial Pr sure Regulation. The nervoua regulators of ar rial pressure, though acting very rapidly and po erfully to correct acute abnormalities of arter

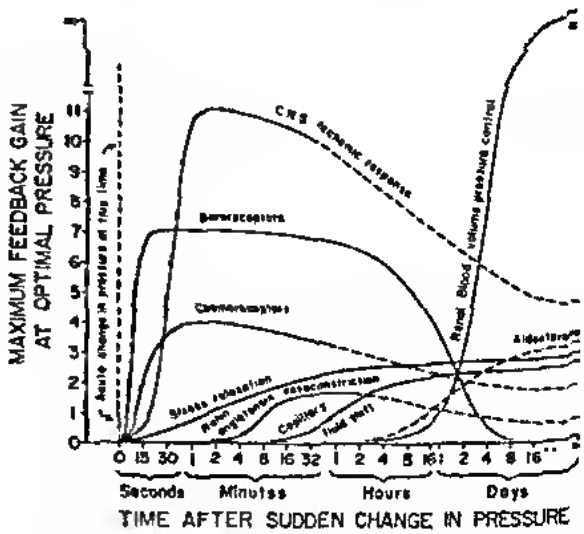


Figure 21–3. Approximate potency of various arterial press control mechanisms at different time intervals after the onset a disturbance to the arterial pressure Note especially the infi gain of the renal–body fluid pressure control mechanism t occurs after a few days' time. (Reprinted from Guyton: Arte Pressure and Hypertension. Philadelphia, W B Saunders Co pany, 1980)

pressure, generally lose their power to control arterial pressure after a few hours to a few days because the nervous pressure receptors "adapt"; that is, they lose their responsiveness. Therefore, except under unusual circumstances, the nervous mechanisms for arterial pressure control almost certainly do not play a major role in long-term regulation of arterial pressure. Long-term regulation, instead, is vested mainly in a renal-body fluid-pressure control mechanism which is illustrated by the curve farthest to the right in Figure 21–3. As we shall discuss in the following chapter, this mechanism involves control of blood volume with its consequent effects on arterial pressure. Part of this mechanism involves, also, control of kidney function by several different hormonal systems, including especially the renin-angiotensin system and the hormone aldosterone secreted by the adrenal cortex.

Basically, the renal-body fluid-pressure control mechanism works as follows: When the arterial pressure falls, this fall in pressure has a direct effect on kidney excretion that causes retention of fluid in the body and therefore increases the blood volume. This, in turn, eventually will raise the blood pressure back to normal. The special value of this mechanism is that it will not stop increasing the blood volume until the pressure rises all the way back to normal. This characteristic of the mechanism makes it an extremely powerful one, as will be described in the following chapter.

RAPIDLY ACTING NERVOUS MECHANISMS FOR ARTERIAL PRESSURE CONTROL

THE ARTERIAL BARORECEPTOR CONTROL SYSTEM—BARORECEPTOR REFLEXES

By far the best known of the mechanisms for arterial pressure control is the *baroreceptor reflex*. Basically, this reflex is initiated by stretch receptors, called either *baroreceptors* or *pressoreceptors*, located in the walls of the large systemic arteries. A rise in pressure stretches the baroreceptors and causes them to transmit signals into the central nervous system, and "feedback" signals are then sent back through the autonomic nervous system to the circulation to reduce arterial pressure downward toward the normal level.

Physiologic Anatomy of the Baroreceptors, and Their Innervation. Baroreceptors are spray-type nerve endings lying in the walls of the arteries; they are stimulated when stretched. A few baroreceptors are located in the wall of almost every large artery of the thoracic and neck regions; but, as illustrated in Figure 21–4, baroreceptors are extremely abundant, in (1) the wall of each internal carotid artery slightly above the carotid bifurcation, an area known as the *carotid sinus*, and (2) the wall of the *aortic arch*.

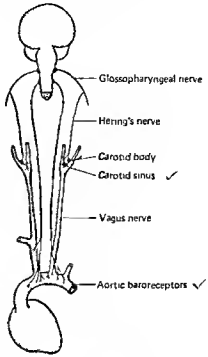


Figure 21–4. The baroreceptor system

Figure 21–4 also shows that signals are transmitted from each carotid sinus through the very small *Hering's nerve* to the glossopharyngeal nerve and thence to the *tractus solitarius* in the medullary area of the brain stem. Signals from the arch of the aorta are transmitted through the vagus nerves also into this area of the medulla. Hering's nerve is especially important in physiologic experiments because baroreceptor impulses can be recorded from it with ease.

Response of the Baroreceptors to Pressure. Figure 21–5 illustrates the effect of different arterial pressures on the rate of impulse transmission in a Hering's nerve. Note that the carotid

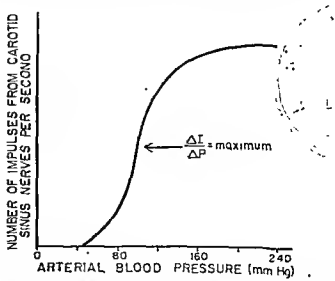


Figure 21–5. Response of the baroreceptors at different levels of arterial pressure.

sinus baroreceptors are not stimulated at all by pressures between 0 and 60 mm Hg, but above 60 mm Hg they respond progressively more and more rapidly and reach a maximum at about 180 mm Hg. The responses of the aortic baroreceptors are similar to those of the carotid receptors, except that they respond, in general, at pressure levels about 30 mm Hg higher.

Note especially that the increase in number of impulses for each unit change in arterial pressure, expressed as $\Delta I/\Delta P$ in the figure, is greatest at a pressure level near the normal mean arterial pressure. That is, in the normal operating range of arterial pressure, even a slight change in pressure causes strong autonomic reflexes to readjust the arterial pressure back toward normal. Thus, the baroreceptor feedback mechanism functions most effectively in the very pressure range where it is most needed.

The baroreceptors respond extremely rapidly to changes in arterial pressure; in fact, the rate of impulse firing increases during systole and decreases again during diastole. Furthermore, the baroreceptors *respond much more to a rapidly changing pressure* than to a stationary pressure. That is, if the mean arterial pressure is 150 mm Hg but at that moment is rising rapidly, the rate of impulse transmission may be as much as twice that when the pressure is stationary at 150 mm Hg. On the other hand, if the pressure is falling, the rate might be as little as one quarter that for the stationary pressure.

The Reflex Initiated by the Baroreceptors. After the baroreceptor signals have entered the tractus solitarius of the medulla, secondary signals *inhibit the vasoconstrictor center* of the medulla and *excite the vagal center*. The net effects are (1) *vasodilatation* throughout the peripheral circulatory system, and (2) *decreased heart rate* and *strength of heart contraction*. Therefore, excitation of the baroreceptors by pressure in the arteries reflexly *causes the arterial pressure to decrease* because of both a decrease in peripheral resistance and a decrease in cardiac output. Conversely, low pressure has opposite effects, reflexly causing the pressure to rise back toward normal.

Figure 21–6 illustrates a typical reflex change in arterial pressure caused by occluding the common carotid arteries. This reduces the carotid sinus pressure, as a result, the baroreceptors become inactive and lose their inhibitory effect on the vasomotor center. The vasomotor center then becomes much more active than usual, causing the arterial pressure to rise and to remain elevated during the ten minutes that the carotids are occluded. Removal of the occlusion allows the pressure to fall immediately to slightly below normal as a momentary overcompensation and then to return to normal in another minute or so.

Function of the Baroreceptors During Changes in Body Posture. The ability of the baroreceptors to maintain relatively constant arterial pressure is extremely important when a person sits or stands after having been lying down. Immediately upon standing, the arterial pressure in the head and upper part of the body obviously tends to fall, and marked reduction of this pressure can cause loss of consciousness. Fortunately, however, the falling pressure at the baroreceptors elicts an immediate reflex, resulting in strong sympathetic discharge throughout the body, and this minimizes the decrease in pressure in the head and upper body.

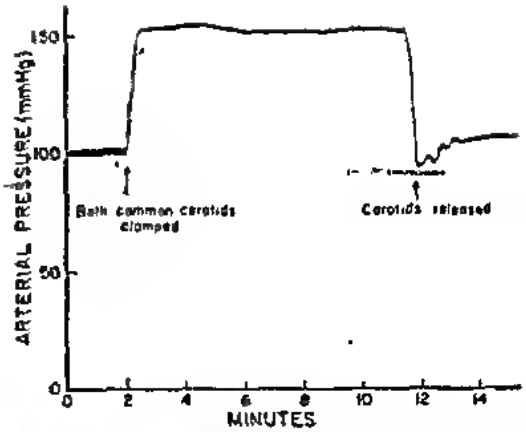


Figure 21–6. Typical carotid sinus reflex effect on arterial pressure caused by clamping both common carotids (after the two vagus nerves have been cut)

The "Buffer" Function of the Baroreceptor Control System. Because the baroreceptor system opposes increases and decreases in arterial pressure, it is often called a *pressure buffer system*, and the nerves from the baroreceptors are called *buffer nerves*.

Figure 21–7 illustrates the importance of this buffer function of the baroreceptors. The upper record in this figure shows an arterial pressure recording for two hours from a normal dog and the lower record from a dog whose baroreceptor nerves from both the carotid sinuses and the aorta had previously been removed. Note the extreme variability of pressure in the denervated dog caused by simple events of the day such as lying down, standing, excitement, eating, defecation, noises, and so forth.

Figure 21–8 illustrates the frequency distributions of the arterial pressures recorded for a full 24-hour day in both the normal dog and the denervated dog. Note that when the baroreceptors were normal, the arterial pressure remained throughout the day within a narrow range between 85 and 115 mm Hg—indeed, during most of the day at almost exactly 100 mm Hg. On the other hand, after denervation of the baroreceptors, the frequency distribution curve became the broad, low curve of the figure, showing that the pressure range increased 2.5-fold, frequently falling to as low as 50 mm Hg or rising to over 160 mm Hg.

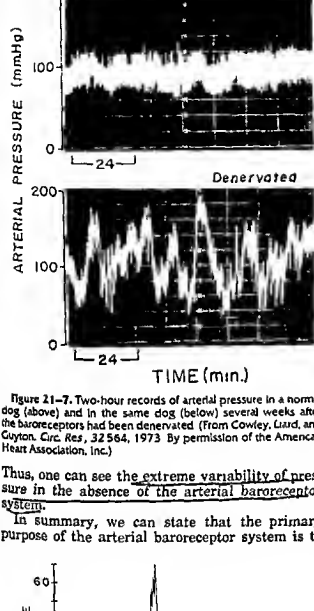


Figure 21–7. Two-hour records of arterial pressure in a normal dog (above) and in the same dog (below) several weeks after the baroreceptors had been denervated (from Cowley, Liard, and Guyton. *Circ. Res.*, 32:564, 1973. By permission of the American Heart Association, Inc.)

Thus, one can see the extreme variability of pressure in the absence of the arterial baroreceptor system.

In summary, we can state that the primary purpose of the arterial baroreceptor system is to reduce the daily variation in arterial pressure to about one half to one third that which would occur were the baroreceptor system not present.

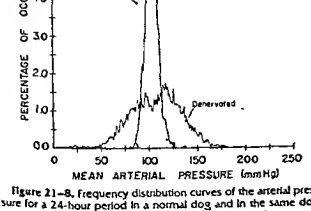


Figure 21–8. Frequency distribution curves of the arterial pressure for a 24-hour period in a normal dog and in the same dog several weeks after the baroreceptors had been denervated. (From Cowley, Liard, and Guyton. *Circ. Res.*, 32:564, 1973. By permission of the American Heart Association, Inc.)

Unimportance of the Baroreceptor System for Long-Term Regulation of Arterial Pressure—Adaptation of the Baroreceptors. The baroreceptor control system is probably of little or no importance in long-term regulation of arterial pressure for a very simple reason: The baroreceptors themselves adapt in one to two days to whatever pressure level they are exposed to. That is, if the pressure rises from the normal value of 100 mm Hg to 200 mm Hg, extreme numbers of baroreceptor impulses are at first transmitted. During the next few seconds, the rate of firing diminishes considerably; then it diminishes very slowly during the next one to two days, at the end of which time the rate will have returned essentially to the normal level despite the fact that the arterial pressure remains 200 mm Hg. Conversely, when the arterial pressure falls to a very low value, the baroreceptors at first transmit no impulses at all, but gradually over a day or days the rate of baroreceptor firing returns again to the original control level.

This adaptation of the baroreceptors obviously prevents the baroreceptors reflex from functioning as a control system to buffer arterial pressure changes that last longer than a few days at a time. In fact, referring again to Figures 21–7 and 21–8, one can see that the average arterial pressure over any prolonged period of time is almost exactly the same whether the baroreceptors are present or not. This illustrates the *unimportance of the baroreceptor system for long-term regulation of the arterial pressure* even though it is a potent mechanism for preventing the rapid changes of arterial pressure that occur moment by moment or hour by hour. Instead, prolonged regulation of arterial pressure requires other control systems, principally the renal-body fluid-pressure control system (along with its associated hormonal mechanisms), to be discussed in the following chapter.

The Carotid Sinus Syndrome. Strong pressure on the neck over the bifurcations of the carotids in the human being can excite the baroreceptors of the carotid sinuses, causing the arterial pressure to fall as much as 20 mm Hg in the normal person. In some older persons, particularly after calcified arteriosclerotic plaques have developed in the carotid arteries, pressure on the carotid sinuses often causes such strong responses that the heart stops completely, or at least the arterial pressure falls drastically. Even tight collars can cause the arterial pressure to fall low enough to cause fainting in these persons, an effect called *carotid sinus syncope*. Fortunately, when the reflex causes the heart to stop, the ventricles usually "escape" from the vagal inhibition in approximately 7 to 10 seconds and begin to beat with their own intrinsic rhythm. However, occasionally the ventricles fail to escape, and the patient dies of cardiac arrest. Treatment of the condition consists of surgically stripping the nerves from the carotid arteries above and below the bifurcations.

CONTROL OF ARTERIAL PRESSURE BY THE CAROTID AND AORTIC CHEMORECEPTORS—EFFECT OF OXYGEN LACK ON ARTERIAL PRESSURE

The pressure in the aorta and in the carotid arteries indirectly controls arterial pressure in another way besides through the baroreceptor reflexes It achieves this by stimulating the *chemoreceptors* when the pressure falls too low

The chemoreceptors are chemosensitive cells located in several small organs 1 to 2 mm in size: two *carotid bodies*, one of which lies in the bifurcation of each common carotid artery, and several *aortic bodies* adjacent to the aorta. The chemoreceptors excite nerve fibers that pass along with the baroreceptor fibers through Hering's nerves and the vagus nerves into the vasomotor center.

Each carotid or aortic body is supplied with an abundant blood flow through a small nutrient artery so that the chemoreceptors are always in close contact with the arterial blood Whenever the arterial pressure falls below a critical level, the chemoreceptors become stimulated because of diminished blood flow to the bodies and therefore diminished availability of oxygen and excess carbon dioxide and hydrogen ions that are not removed by the slow flow of blood.

The signals from the chemoreceptors are transmitted into the vasomotor center to excite the vasomotor center, and this elevates the arterial pressure Obviously, this reflex helps to return the arterial pressure back toward the normal level whenever it falls too low. However, it is not a powerful arterial pressure controller in the normal arterial pressure range, because it does not respond strongly until the arterial pressure falls below 80 mm Hg

The chemoreceptor mechanism also increases the arterial pressure whenever the concentration of oxygen in the arterial blood falls below normal or the concentration of carbon dioxide or hydrogen ions rises above normal The increased pressure helps the circulation deliver increased quantities of oxygen to the tissues and remove excess carbon dioxide and hydrogen ions from the tissues

The chemoreceptors will be discussed in much more detail in Chapter 42 in relation to respiratory control, in which they play an even more important role than in pressure control.

ATRIAL AND PULMONARY ARTERY REFLEXES THAT HELP REGULATE ARTERIAL PRESSURE AND OTHER CIRCULATORY FACTORS

Both the atria and the pulmonary arteries have stretch receptors, called *low pressure receptors*, in their walls similar to the baroreceptor stretch receptors of the large systemic arteries When these low pressure receptors are intact, the arterial pressure also changes less in response to changes in blood volume than when they are not present. To give an example, if 300 ml of blood are infused into a dog with all the receptors intact, the arterial pressure will rise only 15 mm Hg. With the arterial baroreceptors denervated, the pressure will rise 50 mm Hg If the low pressure receptors are also denervated, the pressure will rise 120 mm Hg.

Thus, one can see that even though the low pressure receptors in the pulmonary artery and in the cannot detect the systemic arterial pressure, these ceptors, nevertheless, do detect the simultaneous increase in pressure in the low pressure areas of circulation caused by the increase in volume, and they elicit reflexes parallel to the baroreceptor reflexes to make the total reflex system much more potent for control of arterial pressure.

The receptors in both the pulmonary arteries and in the atria operate in an almost identical way as the baroreceptors from the systemic arteries. That is, they inhibit the vasomotor center and thus reflexly decrease the arterial pressure In addition, some atrial receptors play a special role in controlling blood volume, as follows

Atrial Reflexes to the Kidneys—The Volume Reflex. Stretch of the atria also causes reflex dilatation of the afferent arterioles in the kidneys, which is the same reflex effect that occurs in other peripheral arterioles *but is unusually potent in the kidneys. Also, signals are* transmitted simultaneously to the hypothalamus to decrease the secretion of antidiuretic hormone, thereby indirectly affecting kidney function. The decreased afferent arteriolar resistance causes the glomerular capillary pressure to rise, with resultant increase in filtration of fluid into the kidney tubules. The diminution of antidiuretic hormone diminishes the reabsorption of water from the tubules The combination of these two effects therefore causes rapid loss of fluid into the urine, which also serves as a powerful means to return the blood volume back toward normal.

Obviously, all these mechanisms that tend to return the blood volume back toward normal following a volume overload act indirectly as pressure controllers as well as volume controllers, because excess volume drives the heart to greater cardiac output and leads therefore to greater arterial pressure. This volume reflex mechanism will be discussed in more detail in Chapter 38, along with discussion of mechanisms of volume control.

Atrial Reflex Control of Heart Rate (the Bainbridge Reflex). An increase in atrial pressure also causes an increase in heart rate, sometimes increasing the heart rate as much as 75 per cent Part of this increase in heart rate is caused by the direct effect of the increased atrial volume to stretch the S-A node; it was pointed out in Chapter 14 that such direct stretch can increase the heart rate as much as 15 per cent. An additional 40 to 60 per cent increase in rate is caused by a reflex called the *Bainbridge reflex*. The stretch receptors of the atria that elicit the Bainbridge reflex transmit their afferent signals through the vagus nerves to the medulla of the brain. Then, efferent signals are transmitted back through both the vagal and sympathetic nerves to increase the heart's rate and presumably also strength of contraction. Thus, this reflex helps prevent damming of blood in the veins, the atria, and the pulmonary circulation. This reflex obviously has a different purpose from that of controlling arterial pressure and is actually detrimental to pressure control for short periods of time.

THE CNS ISCHEMIC RESPONSE—CONTROL OF ARTERIAL PRESSURE BY THE VASOMOTOR CENTER IN RESPONSE TO DIMINISHED BRAIN BLOOD FLOW

Normally, most nervous control of blood pressure is achieved by reflexes originating in the baroreceptors

he chemoreceptors, and the low pressure receptors, all f which are located in the peripheral circulation outside he brain. However, when blood flow to the vasomotor enter in the lower brain stem becomes decreased nough to cause nutritional deficiency, a condition called t

the neurons in the vasomotor center itself espond directly to the ischemia and become strongly xcited. When this occurs, the systemic arterial pressure ften rises to a level as high as the heart can possibly *ump. This effect is believed to be caused by failure of* he slowly flowing blood to carry carbon dioxide away rom the vasomotor center; the local concentration of arbon dioxide then increases greatly and has an exremely potent effect in stimulating the sympathetic nervous system. It is possible that other factors, such as he buildup of lactic acid and other acidic substances, lso contribute to the marked stimulation of the vaso*center and to the elevation in pressure.* This ori pressure elevation in response to cerebral isehemia is known as the *central nervous system ischemic response or simply CNS ischemic response.*

The magnitude of the ischemic effect on vasomotor activity is tremendous; it can elevate the mean arterial pressure for as long as ten minutes sometimes to as high as 270 mm Hg. *The degree of sympathetic vasoconstriction caused by intense cerebral ischemia is often so great that some of the peripheral vessels become totally or almost totally occluded.* The kidneys, for instance, will entirely cease their production of urine because of arteriolar constriction in response to the sympathetic discharge. Therefore, *the CNS ischemic response is one of the most powerful of all the activations of the sympathetic vasoconstrictor system.*

Importance of the CNS Ischemic Response as a Regulator of Arterial Pressure. Despite the extremely powerful nature of the CNS ischemic response, it does not become very active until the arterial pressure falls far below normal, down to 60 mm Hg and below, reaching its greatest degree of stimulation at a pressure of 15 to 20 mm Hg. Therefore, it is not one of the mechanisms for regulating normal arterial pressure. Instead, it operates principally as an *emergency arterial pressure control system that acts rapidly and extremely powerfully to prevent further decrease in arterial pressure whenever blood flow to the brain decreases dangerously close to the lethal level.* It is sometimes called the "last ditch stand" pressure control mechanism.

The Cushing Reaction. The so-called Cushing reaction is a special type of CNS ischemic response that results from increased pressure in the cranial vault. For instance, when the cerebrospinal fluid pressure rises to equal the arterial pressure, it compresses the arteries in the brain and cuts off the blood supply to the brain. Obviously, this initiates a CNS ischemic response, which causes the arterial pressure to rise. When the arterial pressure has risen to a level higher than the cerebrospinal fluid pressure, blood flows once again into the vessels of the brain to relieve the ischemia. Ordinarily, the blood pressure comes *to a new equilibrium level* slightly higher than the cerebrospinal fluid pressure, thus allowing blood to continue flowing to the brain at all times. A typical Cushing reaction is illustrated in Figure 21–9.

Depressant Effect of Extreme Ischemia on the Vasomotor Center. If cerebral ischemia becomes so severe that *maximum rise in mean arterial pressure* still cannot relieve the ischemia, the neuronal cells begin to suffer metabolically, and within 3 to 10 minutes they become totally inactive. The arterial pressure then falls to about 40 to 50 mm Hg, which is the level to which the pressure falls when the vasomotor center loses all its control of the circulation so that all tonic vasoconstrictor activity is lost. Therefore, it is fortunate that the ischemic response is extremely powerful so that arterial pressure can usually rise high enough to correct brain ischemia before it causes nutritional depression and death of the neuronal cells

SPECIAL FEATURES OF NERVOUS CONTROL OF ARTERIAL PRESSURE

PARTICIPATION OF THE VEINS IN NERVOUS REGULATION OF CARDIAC OUTPUT AND ARTERIAL PRESSURE

Thus far, we have discussed primarily the ability of the nervous system to regulate arterial pressure by altering arterial resistance. However, much of the regulatory effect of the nervous system is carried out through sympathetic vasoconstrictor fibers to the veins. Indeed, the veins constrict in response to even weaker sympathetic stimuli than do the arterioles and arteries.

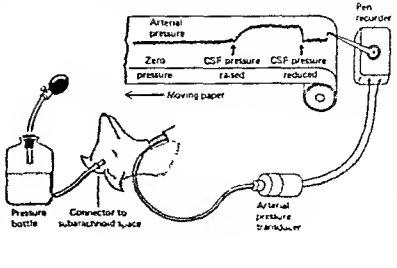


Figure 21–9. The Cushing reaction, showing a rise in arterial pressure resulting from increased cerebrospinal fluid pressure.

Sympathetic Alterations in Venous Capacity and Cardiac Output. The veins offer relatively little resistance to blood flow in comparison with the arterioles and arteries. Therefore, sympathetic constriction of the veins does not significantly change the overall total peripheral resistance. Instead, the important effect of sympathetic stimulation of the veins is a *decrease in their capacity*. This means that the veins then hold less blood at any given venous pressure, which causes translocation of blood out of the systemic veins into the heart, lungs, and systemic arteries. The distension of the heart in turn causes the heart to pump with increasing effectiveness in accordance with the Frank-Starling law of the heart, as discussed in Chapter 13. Therefore, the net effect of sympathetic stimulation of the veins is to increase the cardiac output, which in turn elevates the arterial pressure.

The veins participate in all of the reflexes and reactions that have been discussed thus far, including the baroreceptor reflex, the CNS ischemic response, the chemoreceptor reflex, and also the atrial reflexes. To comprehend the potency of the venous reaction in some of these nervous mechanisms, consider the following: The mean circulatory filling pressure (mainly determined by the degree of contraction of the veins when the blood volume is normal) rises from the normal value of 7 mm Hg to approximately 10 mm Hg when the baroreceptor reflex is excited fully and rises to approximately 18 mm Hg—enough to increase the pressure forcing blood into the heart by as much as 2.5-fold—when a maximal CNS ischemic response is elicited.

ROLE OF THE SKELETAL NERVES AND SKELETAL MUSCLES IN INCREASING CARDIAC OUTPUT AND ARTERIAL PRESSURE

Though most nervous control of the circulation is effected through the autonomic nervous system, at least two conditions in which the skeletal nerves and muscles also play major roles in circulatory responses are the following:

The Abdominal Compression Reflex. When a baroreceptor or chemoreceptor reflex is elicited, or whenever almost any other factor stimulates the sympathetic vasoconstrictor system, the vasomotor center and other areas of the reticular substance of the brain stem transmit impulses simultaneously through skeletal nerves to skeletal muscles of the body, particularly to the abdominal muscles. This increases the basal tone of these muscles, which compresses all the venous reservoirs of the abdomen, helping translocate blood out of the abdominal vascular reservoirs toward the heart. As a result, increased quantities of blood are made available to the heart to be pumped. This overall response is called the *abdominal compression reflex*. The resulting effect on the circulation is the same as that caused by sympathetic vasoconstrictor impulses when they constrict the veins, namely, an increase in both cardiac output and arterial pressure.

The abdominal compression reflex is probably much more important than has been realized in the past, for it is well known that persons whose skeletal muscles have been paralyzed are considerably more prone to hypotensive episodes than are normal persons.

Increased Mean Circulatory Filling Pressure and Increased Arterial Pressure Caused by Skeletal Muscle Contraction During Exercise. When the skeletal muscles contract during exercise, they compress blood vessels throughout the body. And even anticipation of exercise tightens the muscles, thereby compressing the vessels. The resulting effect is to increase the *mean circulatory filling pressure* from its normal value of about 7 mm Hg to as high as 20 to 30 mm Hg. We shall see in Chapter 23 that this high filling pressure acts to translocate blood from the peripheral vessels into the heart and lungs and therefore to increase the cardiac output. This is an essential effect in helping to cause the five- to sixfold increase in cardiac output that sometimes occurs in severe exercise. And the increase in cardiac output in turn is an essential ingredient in increasing the arterial pressure during exercise, an amount of increase usually in the range of 20 to 60 percent.

RESPIRATORY WAVES IN THE ARTERIAL PRESSURE

With each cycle of respiration, the arterial pressure rises and falls usually 4 to 6 mm Hg in a wavelike fashion, giving rise to so-called *respiratory waves* in the arterial pressure. These result from several different effects, some of which are reflex in nature, as follows:

First, many impulses arising in the respiratory center of the medulla "spill over" into the vasomotor center.

Second, every time a person inspires, the pressure in the thoracic cavity becomes more negative than usual, causing the blood vessels in the chest to expand. This reduces the quantity of blood returning to the left side of the heart and thereby momentarily decreases the cardiac output and arterial pressure.

Third, the pressure changes caused in the thoracic vessels by respiration can excite vascular and atrial stretch receptors.

Though it is difficult to analyze the exact relationship of all these factors in causing the respiratory pressure waves, the net result during normal respiration is usually an increase in arterial pressure during the early part of expiration and a fall in pressure during the remainder of the respiratory cycle. During deep respiration the blood pressure can rise and fall as much as 20 mm Hg with each respiratory cycle.

ARTERIAL PRESSURE VASOMOTOR WAVES—OSCILLATION OF THE PRESSURE REFLEX CONTROL SYSTEMS

Often in recording arterial pressure in an animal, in addition to the small pressure waves caused by the respiration some much larger waves are noted—as great as 20 to 40 mm Hg at times—that rise and fall much more slowly than the respiratory waves. The duration of each cycle varies from 26 seconds in the anesthetized dog to as short as 7 to 10 seconds in the human being. These waves are called *vasomotor waves* or sometimes "Mayer waves." Such records are illustrated in Figure 21–10, showing the cyclical rise and fall in arterial pressure.

The cause of vasomotor waves is usually oscillation of one or more nervous pressure control mechanisms, some of which are the following:

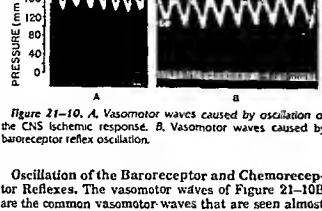


Figure 21–10. A. Vasomotor waves caused by oscillation of the CNS ischemic response. B. Vasomotor waves caused by baroreceptor reflex oscillation.

Oscillation of the Baroreceptor and Chemoreceptor Reflexes. The vasomotor waves of Figure 21–10B are the common vasomotor waves that are seen almost daily in experimental pressure recordings, though usually much less intense than shown in the figure. They are caused mainly by oscillation of the *baroreceptor reflex*. That is, a high pressure excites the baroreceptors; this then inhibits the sympathetic nervous system and lowers the pressure. The decreased pressure reduces the baroreceptor stimulation and allows the vasomotor center to become active once again, elevating the pressure again to a high value. This high pressure then initiates another cycle, and the oscillation continues on and on.

The *chemoreceptor reflex* can also oscillate to give the same type of waves. Usually, this reflex oscillates simultaneously with the baroreceptor reflex. In fact, it probably plays the major role in causing vasomotor waves when the arterial pressure is in the range of 40 to 80 mm Hg, because in this range chemoreceptor control of the circulation becomes powerful while baroreceptor control becomes weak.

Oscillation of the CNS Ischemic Response. The record in Figure 21–10A resulted from oscillation of the CNS ischemic pressure control mechanism. In this experiment the cerebrospinal fluid pressure was raised to 160 mm Hg, which compressed the cerebral vessels and initiated a CNS ischemic response. When the arterial pressure rose above 160 mm Hg, the ischemia was relieved, and the sympathetic nervous system became inactive. As a result, the arterial pressure fell rapidly back to a lower value, causing medullary ischemia once again. The ischemia then initiated another rise in pressure. Then, again the ischemia was relieved, and the pressure fell. This repeated itself cyclically as long as the cerebrospinal fluid pressure remained elevated.

Thus, any pressure control mechanism can oscillate if the intensity of "feedback" is strong enough. The vasomotor waves are of considerable theoretical importance because they show that the nervous reflexes that control arterial pressure obey identically the same principles as those applicable to mechanical and electrical control systems. For instance, if the feedback "gain" is too great in the guiding mechanism of an automatic pilot for an airplane, the plane oscillates from side to side instead of following a straight course.

HORMONAL MECHANISMS FOR RAPID CONTROL OF ARTERIAL PRESSURE

In addition to the rapidly acting nervous mechanisms for control of arterial pressure, there are at least three hormonal mechanisms that also provide either rapid or moderately rapid control of arterial pressure. These mechanisms are

(1) the norepinephrine-epinephrine vasoconstrictor mechanism;

(2) the vasopressin vasoconstrictor mechanism;

(3) and the renin-angiotensin vasoconstrictor mechanism.

THE NOREPINEPHRINE-EPINEPHRINE VASOCONSTRICTOR MECHANISM

In the previous chapter it was pointed out that stimulation of the sympathetic nervous system not only causes direct nervous excitation of the blood vessels and heart but also causes release by the adrenal medullae of epinephrine and norepinephrine into the circulating blood. These two hormones circulate to all parts of the body and cause essentially the same effects in controlling arterial pressure as direct sympathetic stimulation. That is, they excite the heart, they constrict most of the blood vessels, and they constrict the veins.

Therefore, the different reflexes that regulate arterial pressure by exciting the sympathetic nervous system cause the pressure to rise in two ways: by direct circulatory stimulation and by indirect stimulation through the release of epinephrine and norepinephrine into the blood.

Epinephrine and norepinephrine circulate in the blood for one to three minutes before being destroyed, thus maintaining a slightly prolonged excitation of the circulation. Also, these hormones can reach some parts of the circulation that have no sympathetic nervous supply at all, including the very minute vessels such as the metarterioles. And these hormones have especially potent actions on some vascular beds, particularly the skin vasculature.

In general, the epinephrine and norepinephrine system can be considered to be a part of the total sympathetic mechanism for arterial pressure control. Slight differences between the actions of epinephrine and norepinephrine will be discussed in Chapter 57 in relation to the discussion of the entire autonomic nervous system.

ROLE OF VASOPRESSIN IN RAPID CONTROL OF ARTERIAL PRESSURE

When the arterial pressure falls low, either the same nervous signals that activate the sympathetic nervous system or closely related signals cause the hypothalamus to secrete large quantities of vasopressin by way of the posterior pituitary gland—a mechanism that will be discussed in detail in relation to the pituitary hormones in Chapter 75. The vasopressin in turn has a direct vasoconstrictor effect on the blood vessels, thereby increasing both the total peripheral resistance and the mean circulatory filling pressure, raising the arterial pressure back toward normal.

Until recently, most physiologists believed that the amount of vasopressin secreted in low blood pressure states was not sufficient to play a significant role in compensating for the low pressure. However, recent experiments in animals in which the baroreceptor pressure-controlling mechanism has been removed have shown that the circulating amounts of vasopressin found in the blood following hemorrhage can then increase the arterial pressure as much as 35 to 50 mm Hg. In a recent study by Cowley, the vasopressin system was shown to return the blood pressure about 75 per cent of the way toward normal within a few minutes after acute hemorrhage had decreased the arterial pressure to as low as 50 mm Hg.

Therefore, it is almost certain that vasopressin plays a very important role in re-establishing normal arterial pressure when the pressure falls acutely to dangerously low levels.

Vasopressin also plays an indirect role in the long-term control of arterial pressure through its effects on the kidneys in causing decreased excretion of water. Because of this effect, vasopressin is also called *antidiuretic hormone*. When only minute amounts of vasopressin are secreted—in the order of nanograms—the kidney excretion of water decreases to a minimal amount, an effect that helps increase the blood volume any time the arterial pressure remains low for any period of time. The increased blood volume then helps bring the pressure back to normal. Thus, vasopressin not only plays an important role in acute regulation of arterial pressure but also in long-term regulation, which will be discussed in the following chapter.

THE RENIN-ANGIOTENSIN VASOCONSTRICTOR MECHANISM FOR CONTROL OF ARTERIAL PRESSURE

The hormone *angiotensin II* is one of the most potent vasoconstrictors known. It and vasopressin vie with each other for this distinction. Whenever the arterial pressure falls very low, large quantities of angiotensin II appear in the circulation. This results from a special mechanism involving the kidneys and the release of the enzyme *renin* from the kidneys when the arterial pressure falls too low.

The overall schema for formation of angiotensin and the effect of angiotensin II to increase arterial pressure are illustrated in Figure 21–11. When blood flow through the kidneys is decreased, the *juxtaglomerular cells* (cells located in the walls of the afferent arterioles immediately proximal to the glomeruli) secrete renin into the blood. In addition, sympathetic nerve signals directly to the juxtaglomerular cells can also cause these cells to release renin whenever the sympathetic nervous system becomes activated. Renin itself is an enzyme that splits the end off one of the plasma

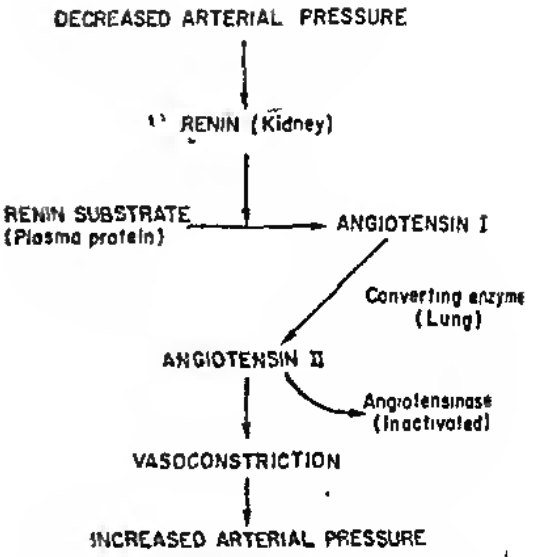


Figure 21–11. The renin-angiotensin-vasoconstrictor mechanism for arterial pressure control.

proteins, called *renin substrate* (or *angiotensinogen*), to release a decapeptide, *angiotensin I*. The renin persists in the blood for 30 minutes to a hour and continues to cause formation of angiotensin I during this entire time. Within a few seconds after formation of the angiotensin I, two additional amino acids are split from it to form the octapeptide *angiotensin II*. This conversion occurs almost entirely in the small vessels of the lungs, catalyzed by an enzyme called *converting enzyme* that is present in the walls of these vessels. Angiotensin II persists in the blood for a minute or so but is rapidly inactivated by a number of different blood and tissue enzymes collectively called *angiotensinase*.

During its persistence in the blood, angiotensin II has several effects that can elevate arterial pressure. One of these effects occurs very rapidly: vasoconstriction, especially of the arterioles and to a lesser extent of the veins. Constriction of the arterioles increases the peripheral resistance and thereby raises the arterial pressure back toward normal, as illustrated at the bottom of the scheme in Figure 21–11. Also, mild constriction of the veins increases the mean circulatory filling pressure, sometimes as much as 20 per cent, and this promotes increased tendency for venous return of blood to the heart, thereby helping the heart pump against the extra pressure load.

The other effects of angiotensin are mainly related to the body fluid volumes: (1) angiotensin has a direct effect on the kidneys to cause decreased excretion of both salt and water; and (2) angiotensin stimulates the secretion of aldosterone by the adrenal cortex, and this hormone in turn also acts on the kidneys to cause decreased excretion of both salt and water. Both these effects tend to elevate the blood volume—an important factor in long-term regulation of arterial pressure, as will be discussed in the following chapter.

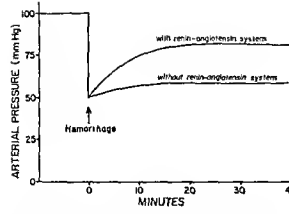


Figure 21–12. Pressure-compensating effect of the renin-angiotensin-vasoconstrictor system following severe hemorrhage (Drawn from experiments by Dr. Royce Brough.)

Rapidity of Action and Pressure Controlling Power of the Renin-Angiotensin Vasoconstrictor System. Figure 21–12 illustrates a typical experiment showing the effect of hemorrhage on the arterial pressure under two separate conditions: (1) with the renin-angiotensin system functioning, and (2) without the system functioning (the system was interrupted by a renin-blocking antibody). Note that following hemorrhage, which causes an acute fall of the arterial pressure to 50 mm Hg, the arterial pressure rose back to 83 mm Hg when the renin-angiotensin system was functional. On the other hand, it rose to only 60 mm Hg when the renin-angiotensin system was blocked. This illustrates that the renin-angiotensin system is powerful enough to return the arterial pressure at least halfway back to normal following severe hemorrhage. Therefore, it can sometimes be of lifesaving service to the body, especially in circulatory shock.

Note that the renin-angiotensin vasoconstrictor system requires approximately 20 minutes to become fully active. Therefore, it is far slower to act than are the nervous reflexes and the norepinephrine-epinephrine system; and it also acts only about one half as rapidly as the vasopressin system. However, it also has a correspondingly longer duration of action.

There is reason to believe that the renin-angiotensin vasoconstrictor system is much more powerful under some conditions than others. For instance, some patients with diseased kidneys secrete tremendous quantities of renin, and the pressure control action of the system is then likely to be very powerful.

TWO INTRINSIC CIRCULATORY MECHANISMS FOR RAPID ARTERIAL PRESSURE REGULATION

In addition to the nervous and hormonal mechanisms for rapid control of arterial pressure, two intrinsic physical mechanisms of the circulation also help control arterial pressure, usually beginning to act within a few minutes and reaching full function within a few hours. These are (1) the capillary fluid shift mechanism, and (2) the vascular stress-relaxation mechanism.

Capillary Fluid Shift. When the arterial pressure changes, this is usually accompanied by a similar change in capillary pressure. This causes fluid to begin moving across the capillary membrane between the blood and the interstitial fluid compartment. Within a few minutes to an hour a new state of equilibrium usually will be achieved, but in the meantime this shift of fluid will have played a very beneficial role in the control of arterial pressure. For instance, if the arterial pressure rises too high, loss of fluid through the capillaries into the interstitial spaces causes the blood volume to fall and thereby causes return of arterial pressure back toward normal. The gain of this system is about 3, which means that it can return the arterial pressure about three-fourths of the way back toward its normal mean value; but it takes effect much more slowly than do the nervous reflex mechanisms.

Stress-Relaxation. When the arterial pressure falls, the pressure usually also falls in most of the blood storage areas such as in the veins, the liver, the spleen, the lungs, and so forth. Conversely, a rise in arterial pressure is often associated with a rise in pressure in these same storage areas. As pointed out in Chapter 18, a pressure change causes vessels gradually to adapt to a new size, thereby accommodating the amount of blood that is available. This phenomenon is called *stress-relaxation* or *reverse stress-relaxation*. Thus, following massive transfusion, the arterial pressure rises markedly at first, but because of relaxation of the circulation during the next ten minutes to an hour, the arterial pressure returns nearly to normal even though the blood volume may be as great as 30 per cent above normal. Conversely, following prolonged bleeding, the reverse stress-relaxation mechanism can cause the blood vessels gradually to tighten around the amount of blood that is left, thereby again re-establishing nearly normal circulatory dynamics.

Unfortunately, the stress-relaxation mechanism has very definite limits, so that acute changes in blood volume of more than approximately plus 30 per cent or minus 15 per cent cannot be corrected by this mechanism.

REFERENCES

Abbod, F M (ed) Disturbances in Neurogenic Control of the Circulation Baltimore, The Williams & Wilkins Co, 1951

Abbod, F M, and Thames, M D Interaction of cardiovascular reflexes in circulatory control *In* Shepherd, J T, and Abbod, F M (eds.) Handbook of Physiology Sec 2, Vol III. Bethesda, American Physiological Society, 1983, p 675

Antonaccio, M J (ed) Cardiovascular Pharmacology New York, Raven Press, 1984

Bishop, V S, Malliani, A., and Thoren, P Cardiac mechanoreceptors *In* Shepherd, J T, and Abbod, F M (eds) Handbook of Physiology Sec 2, Vol III Bethesda, American Physiological Society, 1983, p 497

Blix, A S, and Folkow, B Cardiovascular adjustments to diving in mammals and birds *In* Shepherd, J T, and Abbod, F M (eds.) Handbook of Physiology Sec 2, Vol III Bethesda, American Physiological Society, 1983, p. 917

Bohr, D F, and Webb, R C. Vascular smooth muscle function and its changes in hypertension *Am J Med*, 77 3, 1984

Brezenoff, H E, and Giuliano, R. Cardiovascular control by cholinergic mechanisms in the central nervous system *Annu Rev Pharmacol Toxicol*, 22 341, 1982

Brough, R. B, *et al*: Quantitative analysis of the acute response to

hemorrhage of the renin-angiotensin vasoconstrictor feedback loop in anesthetized dogs. *Cardiovasc Res*, *9* 722–733, 1975
Brown, A. M: Cardiac reflexes. *In* Berne, R. M., *et al.* (eds.) Handbook of Physiology Sec 2, Vol I. Baltimore, Williams & Wilkins, 1979, p 677
Buhler, F R. (ed.) Alpha-Adrenoceptors in Cardiovascular Regulation. New York, Raven Press, 1982
Burattini, R., and Bergdorff, P. Closed loop barorefleex control of total peripheral resistance in the cat. Identification of gains by aid of a model. *Cardiovasc Res*, *18* 715, 1984
Campese, V M, and DeQuattro, V: Orthostatic hypotension Causes and therapy *In* DeGroot, L. J., *et al.* (eds.) Endocrinology Vol 2 New York, Grune & Stratton, 1979, p 1289
Canessa, M The polymorphism of red cell Na and K transport in essential hypertension. Findings, controversies, and perspectives *Prog Clin Biol Res*, *159* 293, 1984
Coleman, T G., *et al.* Angiotensin and the hemodynamics of chronic salt deprivation. *Am J Physiol*, *229* 167–171, 1975
Coleridge, H M, and Coleridge, J C G Cardiovascular afferents involved in regulation of peripheral vessels *Annu Rev Physiol*, *42* 413, 1980
Coleridge, J C G, and Coleridge H M Chemoreflex regulation of the heart *In* Berne, R M, *et al* (eds.) Handbook of Physiology Sec 2, Vol I Baltimore, Williams & Wilkins, 1979, p 653.
Conway, J Hemodynamic aspects of essential hypertension in humans. *Physiol Rev*, *64* 617, 1984
Cowley, A W, Jr, and Guyton, A C. Baroreceptor reflex contribution in angiotensin II induced hypertension *Circulation*, *50* 61, 1974
Cowley, A W, Jr, *et al* Interaction of vasopressin and the baroreceptor reflex system in the regulation of arterial pressure in the dog *Circ Res*, *34* 505, 1974
Cushing, H Concerning a definite regulatory mechanism of the vasomotor center which controls blood pressure during cerebral compression. *Bull Johns Hopkins Hosp*, *12* 290 1901
Dampney R A *et al* Identification of cardiovascular cell groups in the brain stem *Clin Exp Hypertens* *6* 205 1984
Donald, D E., and Shepherd J T Autonomic regulation of the peripheral circulation *Annu Rev Physiol*, *42* 429, 1980
Downing S E. Baroreceptor regulation of the heart *In* Berne, R M., *et al* (eds.) Handbook of Physiology Sec 2 Vol I Baltimore, Williams & Wilkins, 1979, p 621
Edwards, C R, and Padfield, P L Angiotensin-converting enzyme inhibitors. Past, present, and bright future *Lancet* *1* 30 1985
Folkow, B Physiological aspects of primary hypertension *Physiol Rev*, *62* 347, 1982
Guyton, A C Acute hypertension in dogs with cerebral ischemia. *Am J Physiol* *154* 45 1948
Guyton A C Arterial Pressure and Hypertension Philadelphia. W B Saunders Co 1980
Guyton, A C, and Satterfield, J H Vasomotor waves possibly resulting from CNS ischemic reflex oscillation. *Am J Physiol* *170* 601 1952
Guyton, A C *et al* Method for studying competence of the body's blood pressure regulatory mechanisms and effect of pressoreceptor denervation *Am J Physiol* *164* 360, 1951
Guyton, A C *et al* Synthesis of endocrine control in hypertension *Clin Sci Molec Med*, *51* 319, 1976
Hainsworth R. and Linden R J Reflex control of vascular capacitance. *In* International Review of Physiology Cardiovascular Physiology III. Vol 18 Baltimore, University Park Press 1979 p 67
Hall J W and Guyton A C Changes in renal hemodynamics and renin release caused by increased plasma oncotic pressure *Am J Physiol*, *231* 1550, 1976
Herd, J A., Cardiovascular response to stress in man. *Annu Rev Physiol*, *46* 177, 1984
Hilton, S M., and Spyer, K M Central nervous regulation of vascular resistance *Annu. Rev Physiol*, *42* 399, 1980
Kreye, V A., Direct vasodilators with unknown modes of action The nitro-compounds and hydralazine *J Cardiovasc Pharmacol*, *6*(Suppl 4) S646, 1984
Krieger, E M Time course of baroreceptor resetting in acute hypertension *Am J Physiol*, *218* 484, 1970
Ludbrook, J Reflex control of blood pressure during exercise *Annu Rev Physiol*, *45* 155 1983
Mancia, G, and Mark, A. L: Arterial baroreflexes in humans. *In* Shepherd, J T, and Abboud, F M (eds.) Handbook of Physiology Sec 2, Vol III Bethesda, American Physiological Society, 1983, p 755
Mitchell, J H, and Schmidt R. F Cardiovascular reflex control by afferent fibers from skeletal muscle receptors *In* Shepherd, J T, and Abboud, F M (eds.) Handbook of Physiology Sec. 2, Vol III Bethesda, American Physiological Society, 1983, p 623
Opie, L H (ed.) Calcium Antagonists and Cardiovascular Disease New York, Raven Press, 1984
Randall, W C (ed.) Nervous Control of Cardiovascular Function New York, Oxford University Press, 1984
Regoli, D Neurohumoral regulation of precapillary vessels: The kallikrein-kinin system *J Cardiovasc Pharmacol*, *6*(Suppl 2) S401, 1984
Reid I A, *et al* The renin-angiotensin system *Annu Rev Physiol*, *40* 377, 1978
Rowell, L B Reflex control of regional circulations in humans *J Autonom Nerv Syst*, *11* 101, 1984
Sagawa, K. Baroreflex control of systemic arterial pressure and vascular bed *In* Shepherd, J T, and Abboud, F M (eds.) Handbook of Physiology Sec 2, Vol. III Bethesda, American Physiological Society, 1983, p 453
Sagawa, K, *et al* Quantitation of cerebral ischemic pressor response in dogs. *Am J Physiol*, *200* 1164, 1961
Sagawa K, *et al* Dynamic performance and stability of cerebral ischemic pressor response *Am J Physiol*, *201* 1164, 1961
Sagawa K, *et al* Elicitation of theoretically predicted feedback oscillation in arterial pressure *Am J Physiol*, *203* 141, 1962
Smith, E E., and Guyton A C Center of arterial pressure regulation during rotation of normal and abnormal dogs. *Am J Physiol*, *204* 979, 1963
Stephenson, R B Modification of reflex regulation of blood pressure by behavior *Annu Rev Physiol*, *46* 133, 1984
Stiles G L, *et al* β-Adrenergic receptors Biochemical mechanisms of physiological regulation *Physiol Rev*, *64* 661, 1984
Uther, J B and Guyton A C Cardiovascular regulation following changes in central nervous perfusion pressure in the unanesthetized rabbit. *Aust J Exp Biol Med Sci*, *51* 295, 1973
Vanhoutte, P M., Calcium-entry blockers, vascular smooth muscle and systemic hypertension *Am. J Cardiol*, *55* 17B, 1985
Westfall T C Neuroeffector mechanisms. *Annu Rev Physiol*, *42* 383, 1980
Youmans W B, *et al* The Abdominal Compression Reaction Baltimore, Williams & Wilkins, 1963
Youmans W B., *et al* Control of involuntary activity of abdominal muscles. *Am J Phys Med*, *53* 57, 1974
Ziegler, M G., Postural hypotension. *Annu Rev. Med*, *31* 239, 1980

22

Arterial Pressure Regulation: II. Long-Term Control by the Renal–Body Fluid Mechanism and the Renin-Angiotensin System: Mechanisms of Hypertension

Differences Between Long-Term and Short-Term Mechanisms for Pressure Control

There are three major differences between the short-term and long-term mechanisms for arterial pressure control. First, the short-term mechanisms begin to act within seconds to minutes after the arterial pressure becomes abnormal and therefore can quickly institute measures for bringing the arterial pressure back toward normal. On the other hand, the long-term mechanisms almost invariably are slow to begin acting and therefore have very little to do with arterial pressure control from minute to minute or even from hour to hour. Their function is to control the arterial pressure over a period of days, weeks, months, and years.

A second principal difference is that most short-term pressure control mechanisms lose their capability for pressure control after a few hours or a few days. For instance, almost all, if not all, nervous mechanisms "adapt" within one to three days after they are first activated. That is, the nervous receptors that sense the abnormal pressure lose their responsiveness to the changed pressure and therefore no longer can serve in helping keep the pressure near normal. Similar problems exist for most of the other short-term pressure control mechanisms. By contrast, the effectiveness of most long-term pressure control mechanisms becomes steadily greater rather than less with time.

Third, and most important of all, none of the short-term pressure control mechanisms ever succeed in bringing the arterial pressure all the way back to normal. However, the overall long-term system for pressure control does have this ability to bring the pressure *all the way back to normal*, because one of the long-term pressure control mechanisms, the renal–body fluid mechanism, has a so-called "feedback gain for pressure control" that is infinite, as we will explain in subsequent sections of this chapter.

THE RENAL–BODY FLUID SYSTEM FOR ARTERIAL PRESSURE CONTROL

By far the most important mechanism for long-term control of arterial pressure is the *renal–body fluid system*. Simply expressed, this system works in the following way: When the arterial pressure rises, the rise in pressure directly causes kidney output of water and salt to increase markedly, which is called *pressure diuresis* and *pressure natriuresis*. This in turn causes decreased extracellular fluid volume and decreased blood volume. The decrease in blood volume decreases the pumping by the heart, which returns the arterial pressure back to normal. Conversely, when the pressure falls too low, the output of water and salt by the kidneys decreases far below normal, so that fluid begins to accumulate in the body, the body fluid volumes increase, the cardiac output increases, and the arterial pressure rises.

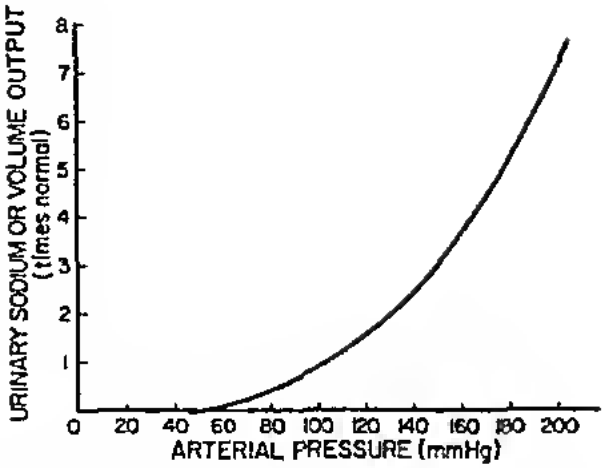


Figure 22–1. A typical renal output curve measured in a perfused isolated kidney, showing both pressure diuresis (excess output of water) and pressure natriuresis (excess output of sodium) when the arterial pressure rises above normal

The basic renal–body fluid system for arterial pressure control is not only the most fundamental mechanism for long-term pressure control, but historically it appeared in the earliest stoges of phylogenetic development of animalhood. For instance, it is fully operative in the very primitive animal, the hagfish This animal constontly drinks sea water that is absorbed into the blood and increases the pressure When the pressure rises too high the kidney simply excretes excessive amounta of urine and relievea the pressure On tho other hand, when the pressure falla too low, the output by the kidneya falls very low, and the pressure bullds up again to its normal range. Through the ages of development of pressure control to the level of human beings, many additional refinements have been added, making pressure control far more exact and far more reliable than in the hagfish

Quantitation of Pressure Diuresis and Pressure Natriuresis As a Basis for Arterial Pressure Control

Figure 22–1 illustrates the average quantitative effects of pressure diuresis and pressure natriuresis as measured in experimentally perfused isolated kidneys. The curve rerorded in this figure is called either a *renal output curve* or a *renal function curve*. Note that at an arterial pressure of 50 mm Hg the urinary output of water and salt is essentially zero. At 100 mm Hg it is normal, and at 200 mm Hg about six to eight times normal. Thus, the effect of arterial pressure on urinary output is a very strong one.

An Experiment Demonstrating the Renal-Body Fluid System for Arterial Pressure Control. Figure 22–2 illustrates an experiment in dogs which all the nervous reflex mechanisms for pressure control were blocked and the arterial pressure was then suddenly elevated by infusing about 300 ml of blood. Note the instantaneous increase in cardiac output to approximately double its normal level and the increase in arterial pressure to 115 mm Hg above its resting level. Also shown, by the middle curve, is the effect of this increased arterial pressure on urinary output. The output increased twelve-fold, and both the cardiac output and the arterial pressure returned to normal during the subsequent hour. Thus, one sees the extreme capability of the kidneys to readjust the blood volume and in so doing to return the arterial pressure back to normal.

Graphical Analysis of Pressure Control by the Renal–Body Fluid Mechanism, Demonstrating Its "Infinite Gain" Feature. Figure 22–3 illustrates a graphical method that can be used for analyzing arterial pressure control by the renal–body fluid system. This analysis is based on two separate curves that intersect each other: (1) the renal output curve for water and salt, which is the same renal output curve as that illustrated in Figure 22–1, and (2) the curve (or line) that represents the level of water and salt intake.

Obviously, over a long period of time the water and salt must equal the intoke. Furthermore, the only place on the graph in Figure 22–3 at which the output equals the intake is where the two curves intersect, which is called the *equilibrium point* Now, let us see what will happen if the arterial pressure becomes some value that is different from that at the equilibrium point:

First, assume that the arterial pressure rises to

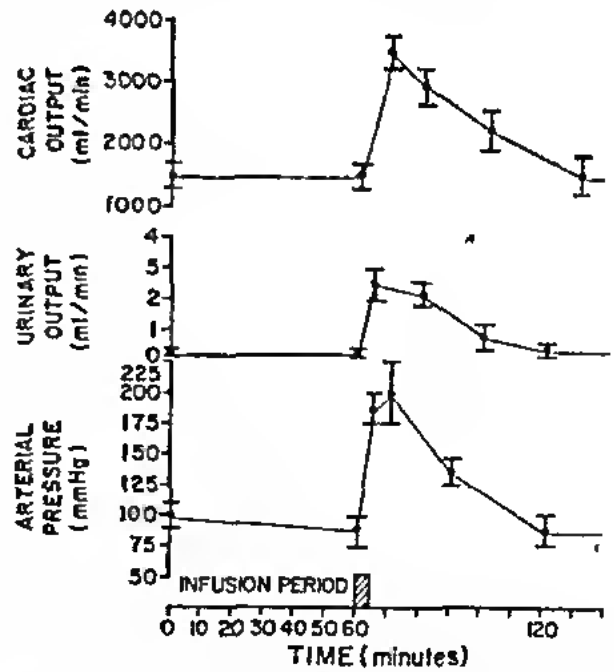


Figure 22–2. Increase in cardiac output, arterial pressure, and urinary output caused by increased blood volume in animals whose nervous pressure control mechanisms had been blocked. This figure shows the return of arterial pressure to normal after about an hour of fluid loss into the urine (Courtesy of Dr William Dobbs)

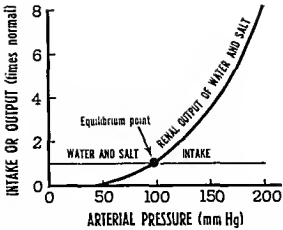


Figure 22–3. Analysis of arterial pressure regulation by equating the renal output curve with the salt and water intake curve. The "equilibrium point" describes the level to which the arterial pressure will be regulated.

150 mm Hg. At this level, the graph shows that renal output of water and salt is about three times as great as the intake. Therefore, the body loses fluid, the blood volume decreases, and the arterial pressure will decrease. Furthermore, this "negative balance" of fluid will not cease until the pressure falls *all the way* back exactly to the equilibrium point. Indeed, even when the arterial pressure is only 1 mm Hg greater than the equilibrium point, there will still be more loss of water and salt than intake, so that the pressure will still continue to fall that last 1 mm Hg *until it returns exactly to the equilibrium point.*

Now, let us see what will happen if the arterial pressure falls below the equilibrium point. This time, the intake of water and salt is greater than the output. Therefore, the body fluid volume increases, the blood volume increases, and the arterial pressure rises until once again it returns *exactly* to the equilibrium point.

This return of the arterial pressure always exactly back to the equilibrium point is the *infinite gain principle* for control of arterial pressure by the renal–body fluid mechanism.

The Two Determinants of the Long-Term Arterial Pressure Level. In Figure 22–3 one can also see that two basic long-term factors determine the long-term arterial pressure level. This can be explained as follows:

As long as the two curves representing (a) renal output of salt and water and (b) salt and water intake remain exactly as they are illustrated in Figure 22–3, the long-term mean arterial pressure level will always readjust exactly to 100 mm Hg, which is the pressure level depicted by the equilibrium point of this figure. Furthermore, there are only two ways in which the pressure of this equilibrium point can be changed from the 100 mm Hg level. One of these is by shifting the renal output curve for salt and water, and the other is by changing the level of the water and salt intake curve. Therefore, expressed very simply, the two primary determinants of the long-term arterial pressure level are

(1) The *degree of shift of the renal output curve* for water and salt along the arterial pressure axis, and

(2) The *level of the water and salt intake curve.*

The operation of these two determinants in the control of arterial pressure is illustrated in Figure 22–4. In Section A of this figure, some abnormality of the kidney has caused the renal output curve to shift 50 mm Hg in the high pressure direction. And note that the equilibrium point has also shifted to 50 mm Hg higher than normal. Therefore, one can state that if the intake of salt and water remains constant but the renal output curve shifts to a new pressure level, so also will the arterial pressure follow to this new pressure level within a few days' time.

Section B of Figure 22–4 illustrates how a change in the level of salt and water intake can change the arterial pressure when the renal output curve remains undisturbed. In this case, the intake has increased four-fold, and the equilibrium point has shifted to a pressure level of 160 mm Hg, 60 mm Hg above the normal level. Conversely, a decrease in the intake level would reduce the arterial pressure.

Therefore, it is *impossible* to change the long-term mean arterial pressure level to a new value without changing one or both of the two basic determinants of the long-term arterial pressure

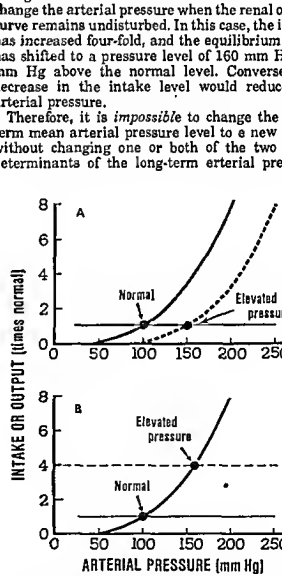


Figure 22–4. Demonstration of the two ways in which the arterial pressure can be increased. *A*, by shifting the renal output curve in the right-hand direction toward a higher pressure level, and *B*, by increasing the intake level of salt and water.

level, either the level of salt and water intake or the degree of shift of the renal function curve along the pressure axis. However, if either of these is changed, one would expect the arterial pressure thereafter to be regulated at a new pressure level.

Failure of Increased Total Peripheral Resistance to Increase the Long-Term Level of Arterial Pressure If Salt Intake and Renal Function Do Not Change

Now is the chance for the student to see whether or not he or she really understands the renal–body fluid mechanism for arterial pressure control. Recalling the basic equation for arterial pressure, *arterial pressure* equals *cardiac output* times *total peripheral resistance*, it is clear that an increase in total peripheral resistance should increase the arterial pressure. And, indeed, when the total peripheral resistance is acutely increased, the arterial pressure does rise immediately. Yet, if the kidneys continue to function normally, the acute rise in arterial pressure will not be maintained. Instead the arterial pressure will return to normal within a day or so. Why?

The answer to this is. Increasing the resistance in the blood vessels everywhere else in the body besides in the kidneys does not change the equilibrium point for blood pressure control (see again Fig. 22–3 and 22–4). Therefore, the kidneys immediately begin to respond to the high arterial pressure with pressure natriuresis and pressure diuresis. Within hours or days, large amounts of salt and water are lost from the body, and this continues until the arterial pressure returns to the pressure level of the equilibrium point.

As proof of this principle, that changes in total peripheral resistance will not affect the long-term level of arterial pressure, carefully study Figure 22–5. This figure shows the cardiac outputs and the arterial pressures in different clinical conditions in which the *long-term* total peripheral resistance is either much less than or much greater than normal, but kidney excretion of salt and water is normal or nearly normal. Note in all the different conditions that the arterial pressure is also exactly normal.

However, note also in Figure 22–5 that a *long-term decrease* in total peripheral resistance causes an exact reciprocal *increase* in cardiac output whereas a long-term increase in total peripheral resistance causes an exact reciprocal *decrease* in cardiac output. We will leave it to the student to figure out how it is that the renal–body fluid mechanism automatically adjusts the cardiac output to the reciprocal level that will balance exactly the changes in total peripheral resistance. Just apply the principle of infinite gain of the renal–body fluid mechanism for pressure control, and the result is inevitable. Until this is clear, one can be sure that he or she still does not understand the renal–body fluid mechanism for arterial pressure control.

(Yet, a word of caution! Many times when the total peripheral resistance increases, this increases the intrarenal vascular resistance at the same time, which alters the function of the kidney and can cause hypertension by shifting the renal function curve to a higher pressure level, in the manner illustrated in Figure 22–4A. We will see an example of this later in this chapter when we discuss hypertension caused by vasoconstrictor mechanisms. But it is the increase in renal resistance that is the culprit, not the increased total peripheral resistance—a very important distinction!)

How Does Increased Fluid Volume Increase the Arterial Pressure? The Role of Autoregulation

The overall mechanism by which increased extracellular volume increases arterial pressure is given in the schema of Figure 22–6. The sequential events are (1) increased extracellular fluid volume, which (2) increases the blood volume, which (3) increases the mean circulatory filling pressure, which (4) increases the venous return of blood to the heart, which (5) increases the cardiac output, which (6) increases the arterial pressure.

Note especially in this schema the two different ways in which an increase in cardiac output can increase the arterial pressure. One of these is (a) the direct effect of increased cardiac output in increasing the pressure, and the other is (b) an indirect effect resulting from local tissue autoregulation of blood flow. This second effect can be explained as follows.

Referring back to Chapter 20, let us recall that

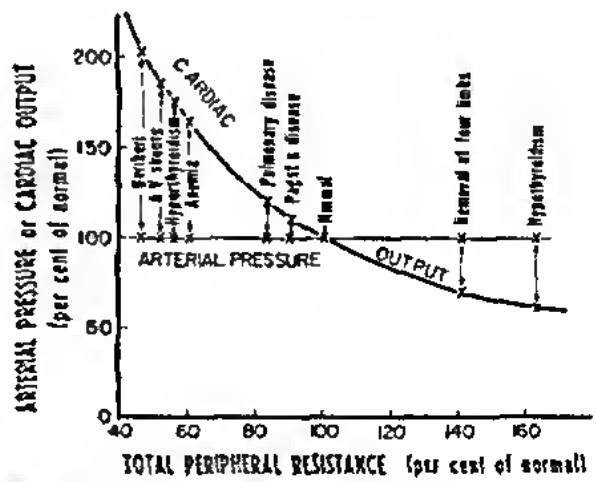


Figure 22–5. Relationships of total peripheral resistance to the long-term levels of arterial pressure and cardiac output in different clinical abnormalities. In these conditions, the kidneys were functioning normally. Note that the changes in total peripheral resistance caused equal and opposite changes in cardiac output but had no effect on the arterial pressure. (Reprinted from Guyton: Arterial Pressure and Hypertension. Philadelphia, W. B. Saunders Company, 1980.)

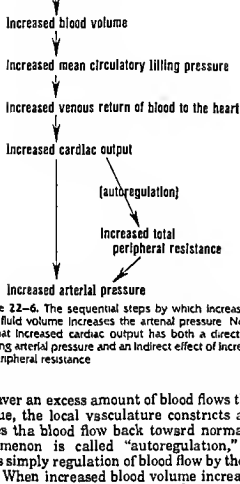


Figure 22–6. The sequential steps by which increased extracellular fluid volume increases the arterial pressure. Note especially that increased cardiac output has both a direct effect of increasing arterial pressure and an indirect effect of increasing the total peripheral resistance.

whenever an excess amount of blood flows through a tissue, the local vasculature constricts and decreases the blood flow back toward normal. This phenomenon is called "autoregulation," which means simply regulation of blood flow by the tissue itself. When increased blood volume increases the cardiac output, the blood flow increases in all the tissues of the body, so that this autoregulation mechanism will constrict the blood vessels all over the body. This, in turn, will also increase the total peripheral resistance.

Finally, since arterial pressure is equal to cardiac output times total peripheral resistance, the secondary increase in total peripheral resistance that results from the autoregulation mechanism greatly increases the rise in arterial pressure.

The Very Slight Excess Fluid Volume That Is Required to Cause Severe Chronic Elevation of the Arterial Pressure. When large quantities of fluid are administered to a person acutely, this has little effect in elevating the arterial pressure. The principal reason for this is that the short-term arterial pressure–regulating mechanisms, especially the nervous feedback mechanisms discussed in the last chapter, oppose the rise in pressure, and the kidneys then have time to excrete the excess fluid before the pressure rises significantly.

However, when this same person is exposed to even slightly increased extracellular fluid volume over a period of days or weeks, the short-term arterial pressure mechanisms—including the nervous mechanisms—lose either all or most of their ability to prevent a rise in arterial pressure after a day or so. Simultaneously, progressive development of autoregulation—especially of "long-term" autoregulation that results from thickening of the vascular musculature—multiplies the effect of cardiac output on the arterial pressure. Combining these two effects, quantitative analyses in animal experiments have demonstrated that very small increases in extracellular fluid volume can increase the arterial pressure perhaps as much as 50 times in the chronic situation as they can acutely. If the student will think about this for a few moments, the beautiful logic of this effect will be seen. It allows a person to change fluid volume rather drastically for short periods of time, as occurs after a meal, without significantly affecting arterial pressure. Yet, over a long period of time, it allows an extremely accurate mechanism for arterial pressure control without necessitating tremendous changes in body fluid volume. In fact, as small as a 3 to 5 per cent change in extracellular fluid volume that lasts for more than a few days—a change that is actually too small to measure under most circumstances—can alter the arterial pressure as much as 20 to 40 mm Hg.

Importance of Salt in the Renal–Body Fluid Schema for Arterial Pressure Regulation

Though the discussions thus far have emphasized the importance of volume in the regulation of arterial pressure, experimental studies have shown that an increase in salt intake is far more likely to increase the arterial pressure than is an increase in water intake. In fact, only under special circumstances does an increase in water intake have a significant effect in elevating the pressure. Therefore, how does one relate salt intake to the volume factor that is so important in the renal–body fluid mechanism for arterial pressure control? The answer to this is the following:

The volume of extracellular fluid is directly related to the amount of accumulation of salt in the body. There are two basic reasons for this:

(1) When there is excess salt in the body the osmolality of the body fluids increases, and this in turn *stimulates the thirst center*, making the person drink extra amounts of water to dilute the extracellular salt to a normal concentration. This obviously increases the extracellular fluid volume.

(2) The increase in osmolality in the extracellular fluid also stimulates the hypothalamic–posterior pituitary gland secretory mechanism to secrete increased quantities of *antidiuretic hormone*. This is discussed fully in Chapter 36. The antidiuretic hormone in turn causes the kidneys to reabsorb greatly increased quantities of water from the urine before it is excreted, thus diminishing the volume of urine while increasing the extracellular fluid volume.

Thus, for these two important reasons, the amount of salt that accumulates in the body :

main determinant of the extracellular fluid volume. And, since only small increases in extracellular fluid can increase the arterial pressure greatly, the accumulation of even a small amount of extra salt in the body can lead to considerable elevation of the arterial pressure.

Yet, we still need to explain why drinking large amounts of water will not also increase the extracellular fluid volume and cause greatly increased pressure. The reason for this is that an excess of water in the extracellular fluid causes exactly opposite effects to those of excess salt on the secretion of antidiuretic hormone by the hypothalamic–posterior pituitary secretory mechanism. That is, it decreases the secretion of this hormone to an extremely low level. Without the antidiuretic hormone, the renal tubular system can reabsorb far less water than normal. Therefore, the volume of urine becomes very great, and the excess water is removed from the extracellular fluid almost as rapidly as it enters the body. This phenomenon is called *water diuresis;* it too will be discussed fully in Chapter 36.

Does Salt Have Other Effects in Increasing Arterial Pressure Besides Its Effect in Increasing Extracellular Fluid Volume? Many research workers have suggested that salt has other important effects in increasing arterial pressure in addition to its effect in increasing the extracellular fluid volume. One of the most persistent suggestions is that excess salt causes increased arteriolar resistance throughout the body, either because of swelling of the walls of the arterioles or because of increased contraction of the vascular smooth muscle. Yet, in recent experiments special procedures have been used to increase the salt in the extracellular fluid to very high concentrations while using artificial means to prevent the extracellular fluid volume from increasing. In none of these experiments did the arterial pressure rise significantly. Therefore, the conclusion is that the role of salt in arterial pressure control almost certainly is mediated either entirely or almost entirely through its propensity for increasing the extracellular fluid volume.

ROLE OF THE RENIN-ANGIOTENSIN SYSTEM AND OF ALDOSTERONE IN LONG-TERM ARTERIAL PRESSURE REGULATION

The Short-Term and the Long-Term Effects of the Renin-Angiotensin System in Pressure Control. In the last chapter the renin-angiotensin system was discussed as one of the important short-term mechanisms for arterial pressure control. Basically, it achieves this short-term control in the following way. When the arterial pressure falls too low, the kidneys secrete renin; the renin then causes the formation of angiotensin in the blood; the angiotensin increases the total peripheral resistance; and the increase in total peripheral resistance increases the arterial pressure back most of the way to normal. This mechanism requires about 20 minutes to became fully effective, and it is based on the acute vasoconstrictor effects of angiotensin.

However, in the previous section of this chapter we pointed out that if renal function is not changed, a change in total peripheral resistance, though having a powerful effect in short-term control of arterial pressure, will not affect the long-term arterial pressure level. Therefore, by what mechanisms can the renin-angiotensin system affect the long-term pressure level?

Fortunately, angiotensin also has two long-term effects that are potently concerned with long-term arterial pressure regulation. These are:

(1) Angiotensin *directly affects the kidneys to cause salt and water retention* in the body.

(2) Angiotensin *causes the adrenal glands to secrete aldosterone,* and aldosterone in turn also directly affects the kidneys to cause salt and water retention.

Therefore, angiotensin can influence the renal-body fluid mechanism for arterial pressure control and therefore elevate the arterial pressure by two separate mechanisms.

The Direct Effect of Angiotensin on the Kidneys in Causing Salt and Water Retention. Renin is formed in the kidneys, and this renin in turn causes moderate amounts of angiotensin also to be formed in the kidneys even before the renin leaves the kidneys and enters the blood. Furthermore, additional angiotensin is formed by the renin once it has entered the circulating blood, and this angiotensin returns to the kidneys by way of the renal arteries. In the kidneys, the angiotensin has several important effects on renal function, the net result of which is to cause the kidneys to excrete far less sodium than normal. These intrarenal functions of angiotensin will be discussed in detail in Chapters 35 and 36. For reasons discussed earlier in this chapter, the renal retention of sodium causes the kidneys to excrete far less water as well, thus increasing the extracellular fluid volume. This, in turn, has a long-term effect on increasing the arterial pressure.

Stimulation of Aldosterone Secretion by Angiotensin, and the Effect of Aldosterone in Increasing Salt and Water Retention by the Kidneys. Angiotensin is one of the most powerful controllers of aldosterone secretion, as we shall discuss in relation to body fluid regulation in Chapter 36 and in relation to adrenal gland function in Chapter 77. Therefore, when the renin-angiotensin system becomes activated, the rate of aldosterone secretion usually increases at the same time. One of the most important functions of aldosterone is to cause marked increase in sodium reabsorption by the kidney tubules, thus increasing the extracellular fluid sodium. This then causes water retention, as already explained, thus increasing the extracellular fluid volume and lead-

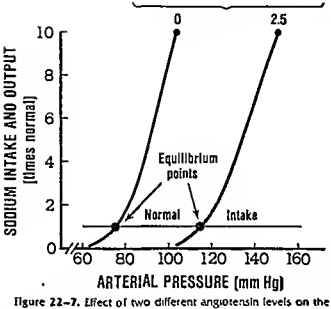


Figure 22–7. Effect of two different angiotensin levels on the renal output curve, showing regulation of the arterial pressure at an "equilibrium point" of 75 mm Hg when the angiotensin level is low and at 115 mm Hg when the angiotensin level is high

ing secondarily to still more long-term elevation of the arterial pressure.

Both the direct effect of angiotensin on the kidney and its effect acting through aldosterone are important in long-term arterial pressure control. However, recent research has suggested that the direct effect on the kidneys is perhaps two to three times as potent as the indirect effect acting through aldosterone—even though the indirect effect is the one most widely known.

Quantitative Analysis of Arterial Pressure Changes Caused by Angiotensin. Figure 22–7 illustrates a quantitative analysis of the type used in Figures 22–3 and 22–4 to analyze arterial pressure control by the renal–body fluid mechanism. This figure shows two separate renal output curves and also illustrates the normal level of sodium intake. The left-hand renal output curve is approximately that measured in dogs whose renin-angiotensin system has been blocked by an appropriate drug. The right-hand curve was measured in dogs infused continuously with angiotensin at a level about 2.5 times the normal rate of angiotensin formation in the blood. Note the obvious shift of the renal output curve toward higher pressure levels under the influence of angiotensin. This shift is caused by both the direct effect of angiotensin on the kidney and the indirect effect acting through aldosterone secretion, as explained above.

Finally, note the two separate equilibrium points, one for zero angiotensin at an arterial pressure level of 75 mm Hg and one for elevated angiotensin at a pressure level of 115 mm Hg.

Therefore, even though the vasoconstrictor effect of angiotensin on all the nonrenal blood vessels of the body cannot cause a long-term increase in arterial pressure, the effects of angiotensin on renal retention of water and salt can have a very powerful effect in causing chronic elevation of the arterial pressure.

Role of the Renin-Angiotensin System in Maintaining a Normal Arterial Pressure Despite Wide Variations in Salt Intake

Probably the most important function of the renin-angiotensin system is to allow a person to eat either a very small or a very large amount of salt without causing great changes in either extracellular fluid volume or arterial pressure. This function is explained by the schema in Figure 22–8 which shows that the direct effect of increased salt intake is to increase the extracellular fluid volume and this, in turn, to increase the arterial pressure. Then, the increased arterial pressure causes increased blood flow through the kidneys and thereby reduces the rate of secretion of renin to a much lower level, this leads sequentially to decreased renal retention of salt and water, return of the extracellular fluid volume to normal, and finally return of the arterial pressure also to normal. Thus, the renin angiotensin system is an automatic feedback mechanism that helps maintain the arterial pressure at or near the normal level even when the salt intake is increased. When the salt intake is decreased far below normal, exactly the opposite effects take place.

To emphasize the importance of this effect of the renin-angiotensin system, the arterial pressure rises no more than a few millimeters of Hg in response to greatly increased salt intake when the system functions normally, but, when it fails to function, the pressure rise is about ten times as great, sometimes rising 50 to 60 mm Hg.

Graphical Analysis of the Combined Role of Sodium Intake and of the Renin-Angiotensin System in Arterial Pressure Control—The "Sodium-Pressure Relationship." Figure 22–9 presents a graphical analysis of the automatic renin-angiotensin feedback mechanism for control of arterial pressure. It shows four separate levels of salt intake, as illustrated by the four separate horizontal lines. To the right of each line is

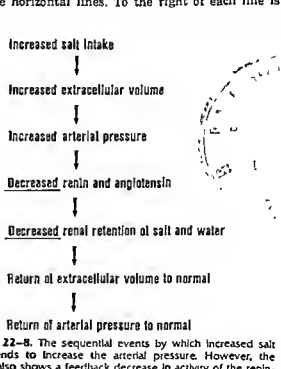


Figure 22–8. The sequential events by which increased salt intake tends to increase the arterial pressure. However, the schema also shows a feedback decrease in activity of the renin-angiotensin system that tends to return the arterial pressure back to the normal level.

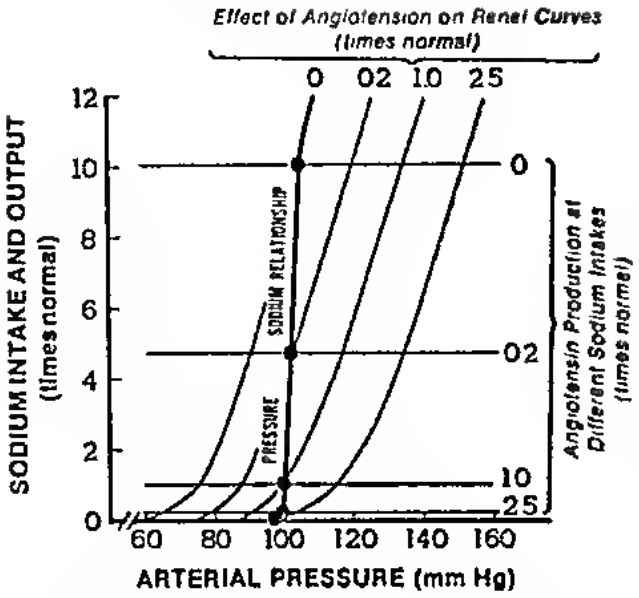


Figure 22–9. The effect of different levels of salt intake on the arterial pressure. The lighter curves are renal output curves. The very dark curve is the pressure-sodium relationship for a person with normal kidneys. See text for explanation.

shown the approximate level of circulating angiotensin that one finds in the blood at each of these salt intake levels. Note that at very low salt intake, only one fifth the normal level, which tends to decrease both the extracellular fluid volume and the arterial pressure, the concentration of angiotensin in the blood increases to about 2.5 times normal. On the other hand, when the sodium intake is 50 times greater, the level of circulating angiotensin becomes essentially zero.

Now, observe the four separate renal output curves (the black curves). These are the renal output curves caused by the four different levels of angiotensin that occur respectively at the different levels of salt intake. Thus, when the level of angiotensin in the blood is high, the renal output curve to the right depicts the relationship between arterial pressure and sodium output. Conversely, when the circulating level of angiotensin is zero, the far left renal output curve depicts the relationship between arterial pressure and sodium output.

Now, finally, match the appropriate levels of sodium intake with the appropriate renal output curves at the different levels of angiotensin and determine the equilibrium points for these four separate pairs of intake and output curves. These equilibrium points are illustrated by the very heavy points on the graph.

As a final maneuver, draw a heavy line connecting the four separate equilibrium points, as is illustrated in the figure. This line represents the relationship between long-term sodium intake and long-term level of arterial pressure, called the *sodium pressure relationship* or the *pressure-sodium relationship*. Note that despite an approximate fifty-fold change in sodium intake, as depicted in the figure, the arterial pressure change is only 4 mm Hg. These are the results that have been observed in animals when the sodium intake has been changed as much as fiftyfold. Furthermore, the same experiment performed in humans has shown that changing the sodium intake from 1/15th normal to 10 times normal—total range of 150-fold—changes the arterial pressure only about 17 mm Hg.

However, another word of caution! This relationship between sodium intake and the long-term arterial pressure level holds true only as long as both the kidneys and the renin-angiotensin system function normally. We shall see later in the chapter that abnormal function of either of these two mechanisms can lead to extreme increases in arterial pressure when the sodium intake is increased.

HYPERTENSION (HIGH BLOOD PRESSURE)

When a person is said to have hypertension (or "high blood pressure"), it is generally meant that his or her mean arterial pressure is greater than the upper range of accepted normality. Usually, a mean arterial pressure of greater than 110 mm Hg under resting conditions is considered to be hypertensive; this level normally occurs when the diastolic blood pressure is greater than 90 mm Hg and the systolic pressure greater than about 135 to 140 mm Hg. In severe hypertension, the mean arterial pressure often rises to as high as 150 to 170 mm Hg, with diastolic pressures as high as 130 to 150 mm Hg and systolic arterial pressures sometimes as great as 250 mm Hg.

The Basic Causes of Hypertension

Now that we have discussed the principles of long-term arterial pressure regulation, we can call on these principles to discuss the mechanisms by which a person can develop hypertension. Experiments in both animals and human beings have demonstrated that hypertension rarely occurs except when the renal function curve becomes abnormal and is shifted to higher pressure levels. That is, if a greater than normal pressure is required to cause the kidneys to excrete salt and water, then the long-term mechanisms for raising the pressure will become progressively more active until the pressure rises to the high level that is required to make the kidneys excrete normally. This can come about either as a result of salt and water retention, or activation of vascular constrictor mechanisms, or both of these. Therefore, for purposes of discussion, hypertension can be classified into two extreme types (but with many gradations between the two):

(1) *Volume-loading hypertension*, and
(2) *Vasoconstrictor hypertension*.

VOLUME-LOADING HYPERTENSION

Volume-loading hypertension occurs when excess extracellular fluid volume accumulates in the body—that is, if all other functions of the circulation are normal.

Let us discuss a typical volume-loading type of hypertension.

Experimental Volume-Loading Hypertension Caused by Reduced Renal Mass Along with Simultaneous Increase in Salt Intake. Figure 22–10 illustrates a typical experiment showing volume-loading hypertension in a group of dogs with 70 per cent of the renal mass removed. At the first circled point on the curve, the two poles of one of the kidneys were removed, and at the second circled point the entire opposite kidney

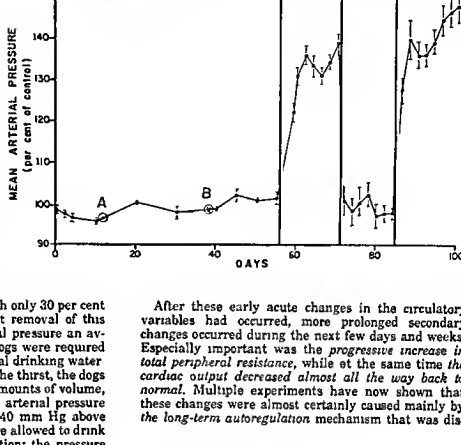


Figure 22–10. Effect on arterial pressure of drinking 0.9 per cent saline solution instead of water in four dogs with 70 per cent of their renal tissue removed. (From Langston, Guyton, Douglas, and Dorsett *Circ. Res.*, 12:508, 1963. By permission of The American Heart Association, Inc.)

was removed, leaving the animals with only 30 per cent of the normal renal mass. Note that removal of this amount of mass increased the arterial pressure an average of only 6 mm Hg. Then, the dogs were required to drink salt solution instead of normal drinking water. Because salt solution fails to quench the thirst, the dogs drank two to four times the normal amounts of volume, and within a few days their average arterial pressure rose to a much higher level, about 40 mm Hg above normal. After two weeks the dogs were allowed to drink tap water again instead of salt solution; the pressure returned to normal within two days. Finally, at the end of the experiment, the dogs were required to drink salt solution again, and this time the pressure rose much more rapidly to an even higher level because the dogs had already learned to tolerate the salt solution and therefore drank much more. Thus, this experiment demonstrates volume-loading hypertension.

The Sequential Effects in the Development of Volume-Loading Hypertension. It is especially instructive to study the sequential changes in circulatory function during the progressive development of volume-loading hypertension. Figure 22–11 illustrates these sequential changes. A week or so prior to the point labeled "0" days, the kidney mass had already been decreased to only 30 per cent of normal. Then at this point, the intake of salt and water was increased to about six times normal. The acute effect was to increase the extracellular fluid volume, the blood volume, and the cardiac output to 20 to 40 per cent above normal. Simultaneously, the arterial pressure began to rise as well, but not nearly so much at first as did the fluid volumes and cardiac output. The reason for this can be discerned by studying the total peripheral resistance curve, which shows an initial *decrease* in total peripheral resistance. This decrease was caused by the baroreceptor mechanism, discussed in the previous chapter, which tried to prevent the rise in pressure. However, after three to four days, the baroreceptors adapted (reset) and no longer opposed the rise in pressure. By this time, the arterial pressure had risen almost to its full height because of the early increase in cardiac output, even though the total peripheral resistance was still almost at the normal level.

After these early acute changes in the circulatory variables had occurred, more prolonged secondary changes occurred during the next few days and weeks. Especially important was the *progressive increase in total peripheral resistance*, while at the same time *the cardiac output decreased almost all the way back to normal*. Multiple experiments have now shown that these changes were almost certainly caused mainly by *the long-term autoregulation* mechanism that was dis-

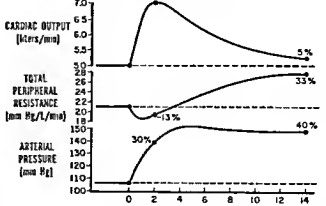


Figure 22–11. Progressive changes in important circulatory system variables during the first few weeks of *volume-loading hypertension*. Note especially the initial rise in cardiac output as the basic cause of the hypertension. Subsequently, the autoregulation mechanism returns the cardiac output almost to normal while at the same time causing a *secondary increase in total peripheral resistance*. (Modified from Guyton: Arterial Pressure and Hypertension. Philadelphia, W. B. Saunders Company, 1980.)

cussed in Chapter 20 and also earlier in this chapter in relation to long-term arterial pressure regulation. That is, after the cardiac output had risen to a high level and had caused hypertension to develop, the excess blood flow through the tissues then caused progressive constriction of the local arterioles, thus returning the cardiac output almost all the way back to normal but simultaneously causing a *secondary increase in total peripheral resistance.*

Note also that the extracellular fluid volume and blood volume returned almost all the way back to normal along with the decrease in cardiac output. This resulted from two factors: First, the increase in arteriolar resistance decreased the capillary pressure, which allowed the fluid in the tissue spaces to be absorbed back into the blood. Second, the elevated arterial pressure now caused the kidneys to excrete the excess volume of fluid that had initially accumulated in the body.

Finally, let us take stock of the final state of the circulation several weeks after the initial onset of volume loading. We find the following effects:

(1) *Hypertension,*

(2) *Marked increase in total peripheral resistance,* and

(3) *Almost complete return of the extracellular fluid volume, blood volume, and cardiac output back to normal.*

Therefore, we can divide volume-loading hypertension into two separate sequential stages.

The first stage in volume-loading hypertension results from increased fluid volumes and increased cardiac output. It is this increase in cardiac output that causes the hypertension.

The second stage in volume-loading hypertension is characterized by high blood pressure and high total peripheral resistance but return of the cardiac output so near to normal that the usual measuring techniques most frequently cannot detect an abnormally elevated output.

It should be especially noted that *the increased total peripheral resistance in volume-loading hypertension occurs after the hypertension has developed and therefore is secondary to the hypertension rather than being the cause of the hypertension.*

Importance of Autoregulation in Maintaining Normal Cardiac Output in Hypertension and in Increasing the Total Peripheral Resistance—But Not in Controlling the Arterial Pressure. In the above example of volume-loading hypertension, almost all the rise in arterial pressure had already occurred before the secondary increase in total peripheral resistance took place as a result of the autoregulation mechanism. Furthermore, even if the autoregulation mechanism had not increased the total peripheral resistance at all, the arterial pressure still would have risen to the same high pressure level. The reason for this is simply that the pressure level is determined by the equilibrium point between the renal output curve and the intake level of salt and water. And the renal-body fluid mechanism will continue to increase the cardiac output to whatever level is required to reach the pressure level of the equilibrium point. If the total peripheral resistance increases, so much the better; under these conditions less increase in cardiac output is required.

Therefore, it is clear that *autoregulation is not a requirement for increasing the arterial pressure.* On the other hand, it plays an extremely important role in returning the cardiac output back to normal once the hypertension has developed (or even *while the hypertension is developing*). Also it is almost certainly the cause of the very great increase in total peripheral resistance that occurs in many types of hypertension when no specific factor can be found that directly constricts the blood vessels.

Volume-Loading Hypertension in Patients Who Have No Kidneys But Are Being Maintained on an Artificial Kidney

When a patient is maintained on an artificial kidney, it is especially important to keep the body fluid volume at a normal level—that is, to remove the appropriate amount of water and salt each time the patient is dialyzed. If this is not done and the extracellular fluid volume is allowed to increase, hypertension will develop in almost exactly the same way as that illustrated in Figure 22–11. That is, the cardiac output increases at first and causes hypertension. Then the autoregulation mechanism returns the cardiac output back toward normal while causing a secondary increase in total peripheral resistance.

Hypertension Caused by Primary Aldosteronism

Another type of volume-loading hypertension is caused by excess aldosterone in the body or occasionally by excesses of other types of steroids. A small tumor in one of the adrenal glands occasionally secretes very large quantities of aldosterone, which is the condition called "primary aldosteronism." As will be discussed in Chapter 35, the aldosterone increases the rate of reabsorption of salt and water by the tubules of the kidneys, thereby reducing the loss of these in the urine but causing an increase in the extracellular fluid volume. Consequently, mild to moderate hypertension occurs, and when salt intake is increased at the same time, the hypertension becomes greater. Furthermore, if the condition persists for months or years, the excess aldosterone seems to cause pathological changes in the kidneys that make the kidneys then retain still more salt and water in addition to that caused by the aldosterone. Therefore, the hypertension can finally become very severe.

Here again, in the early stages of this type of hypertension, the cardiac output is increased, but in the later stages the cardiac output generally returns almost entirely to normal whereas the total peripheral resistance becomes secondarily elevated.

VASOCONSTRICTOR HYPERTENSION

A second type of hypertension that shows sharp contrasts to the volume-loading type is that caused by continuous infusion of a vasoconstrictor agent into the circulation or by excessive secretion of a vasoconstrictor by one of the endocrine glands. The vasoconstrictors that are especially prone to cause hypertension are (1) angiotensin II, (2) norepinephrine, and (3) epinephrine.

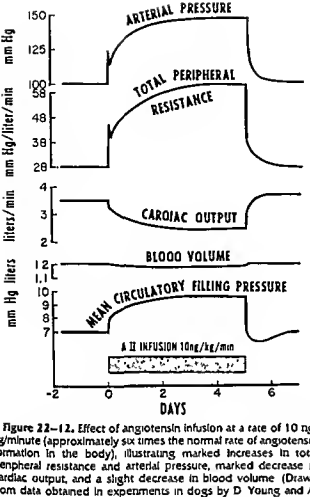


Figure 22–12. Effect of angiotensin infusion at a rate of 10 ng/kg/minute (approximately six times the normal rate of angiotensin formation in the body), illustrating marked increases in total peripheral resistance and arterial pressure, marked decrease in cardiac output, and a slight decrease in blood volume (Drawn from data obtained in experiments in dogs by D Young and A. Cowley)

Hypertension Caused by Continuous Infusion of Angiotensin II or Secretion by a Renin-Secreting Tumor

Figure 22–12 illustrates typical development of vasoconstrictor hypertension caused by infusing angiotensin II at a rate approximately six times the rate of normal angiotensin formation in the body Note especially the marked increases in both total peripheral resistance and arterial pressure but, conversely, a slight decrease in blood volume and a considerable decrease in cardiac output.

All the effects illustrated in Figure 22–12 can be explained by the basic principles of blood pressure regulation that were presented earlier in the chapter:

The increase in total peripheral resistance obviously results from the very potent effect of angiotensin II in constricting the arterioles. It has usually been stated that it is this increase in total peripheral resistance that causes the hypertension However, we shall see in the discussion below that it is the effect of the angiotensin II on kidney excretion of salt and water that is critical in the maintenance of the long-term hypertension.

The slight decrease in blood volume is caused by a very interesting interplay between the increase in total peripheral resistance and the effect of angiotensin II on the kidneys. The increase in total peripheral resistance has an immediate effect of increasing the arterial pressure. On the other hand, the effect of angiotensin on the kidneys is to shift the renal output curve toward a higher pressure level. If the rise in arterial pressure caused by the increase in total peripheral resistance is greater than the shift in the renal output curve, then the initial rise in pressure cannot be sustained by the shift in the kidney output curve. Therefore, this excess pressure causes pressure diuresis and natriuresis, thus causing loss of fluid from the body and decreasing the blood volume slightly. In addition, part of the decrease in blood volume also results from a mild constrictor effect that angiotensin II exerts on the veins.

The *decrease in cardiac output* results almost entirely from the intense arteriolar constriction. In the following chapter we shall discuss the regulation of cardiac output, and we will see that an increase in the resistance to peripheral blood flow has a potent effect on decreasing venous return to the heart, which, in turn, causes the decreased cardiac output because the heart automatically adjusts the cardiac output to equal the venous return.

Thus, it is clear from Figure 22–12 that the hypertension caused by angiotensin-induced vasoconstriction is, like volume-loading hypertension, a high resistance type of hypertension However, there are subtle differences. In volume-loading hypertension, the total peripheral resistance is not increased quite as much as the increase in pressure, and the cardiac output is slightly above normal. On the other hand, in vasoconstrictor hypertension, the total peripheral resistance is increased somewhat in excess of the increase in arterial pressure, while the cardiac output is decreased to below normal.

Vasoconstrictor Hypertension Caused by a Pheochromocytoma. Another type of vasoconstrictor hypertension is that caused by a tumor called a *pheochromocytoma*, which is a tumor of the adrenal medulla that secretes large amounts of *epinephrine* and *norepinephrine* Essentially the same type of vasoconstriction occurs as that in angiotensin hypertension, the total peripheral resistance rises very high, and the blood volume and cardiac output are likely to be low (except when the epinephrine stimulates the body metabolism so much that it sometimes elevates the cardiac output for this reason).

Role of the Renal-Body Fluid Pressure Control Mechanism in Determining the Long-Term Arterial Pressure Level in Vasoconstrictor Hypertension. Let us for a moment suppose that angiotensin were infused into a person at a rate high enough to cause a very large increase in total peripheral resistance and therefore also cause an acute rise in arterial pressure. Let us further suppose that the angiotensin had no effect at all on the renal output curve—that is, the angiotensin did not in any way affect the normal capability of the kidneys to excrete salt and water. In such a theoretical instance as this, the initial marked rise in arterial pressure would cause very intense pressure diuresis and pressure natriuresis. Consequently, the blood volume would decrease rapidly and the arterial pressure eventually return to normal, because it is only at a normal arterial pressure that urinary output would come into balance with the salt and water intake. Therefore, the increase in total peripheral resistance could not sustain a long-term increase in arterial pressure. Instead, to sustain the increased pressure, it is essential that the

kidney output curve be shifted to a high pressure level. Thus, strange as it may seem, even in the vasoconstrictor type of hypertension it is the effect of the vasoconstrictor on kidney output of salt and water that determines the arterial pressure level at which the hypertension stabilizes. If the student understands this concept, it will help tremendously in understanding both blood pressure regulation and the different clinical hypertensive conditions.

TYPES OF HYPERTENSION WITH BOTH VOLUME-LOADING AND VASOCONSTRICTOR COMPONENTS

We shall see in subsequent discussions in this chapter that most clinical types of hypertension are rarely of the pure volume-loading type or the pure vasoconstrictor type but instead exhibit various combinations of both of these. Some of these are the following:

Goldblatt Hypertension

"One-Kidney" Goldblatt Hypertension. When one kidney is removed and a constrictor is placed on the renal artery of the remaining kidney, as illustrated in Figure 22–13, the initial effect is greatly reduced renal arterial pressure, shown by the dashed curve in the figure. However, within a few minutes the systemic arterial pressure begins to rise and continues to rise for several days. The pressure usually rises rapidly for the first hour or so and this is then followed by a slower rise over a period of several days to a much higher pressure level. When the systemic arterial pressure reaches its new stable pressure level, the renal arterial pressure will have returned almost all the way back to normal. The hypertension produced in this way is called *"one-kidney" Goldblatt hypertension* in honor of Dr. Harry Goldblatt, who first studied the important quantitative features of hypertension caused by renal artery constriction.

The early rise in arterial pressure in Goldblatt hypertension is caused by the renin-angiotensin mechanism. Because of the poor blood flow through the kidney after acute application of the constrictor, large quantities of renin are secreted by the kidneys, as illustrated by the lowermost curve in Figure 20–13, and this causes angiotensin to be formed in the blood, as described in the previous chapter; the angiotensin in turn raises the arterial pressure acutely. However, the secretion of renin rises to a peak in a few hours but returns all the way back to normal within five to seven days because the renal arterial pressure by that time has also risen back to normal so that the kidney is no longer ischemic.

The second rise in arterial pressure is caused by fluid retention; within five to seven days the fluid volume has increased enough to raise the arterial pressure to its new sustained level. The quantitative value of this sustained pressure level is determined by the degree of constriction of the renal artery. That is, the aortic pressure must now rise to a much higher than normal level in order to make the renal arterial pressure distal to the constrictor rise high enough to cause normal urinary output.

Note especially that one-kidney Goldblatt hypertension has two phases. The first phase is a vasoconstrictor type of hypertension caused by angiotensin, but this is only transient. The second stage is a volume-loading type of hypertension. However, it is often very difficult to tell that this second stage is true volume-loading hypertension because neither the blood volume nor the cardiac output are significantly elevated. Instead, the total peripheral resistance is increased. To understand this, recall that in pure volume-loading hypertension blood volume and cardiac output are elevated only the first few days during the onset; but after the first few days, volume-loading hypertension is a high resistance hypertension exactly as seen in the second stage of one-kidney Goldblatt hypertension. And the transient increases in blood volume and cardiac output that are normally seen in pure volume-loading hypertension are obscured by the angiotensin vasoconstriction that occurs in Goldblatt hypertension during these early days.

"Two-Kidney" Goldblatt Hypertension. Hypertension also often results when the artery to one kidney is constricted while the artery to the other kidney is still normal. This hypertension results from the following mechanism: The constricted kidney retains salt and water because of decreased renal arterial pressure in this kidney. Also, the "normal" kidney retains salt and water because of renin produced by the ischemic kidney. This renin causes the formation of angiotensin, which circulates to the opposite kidney to cause it also to retain

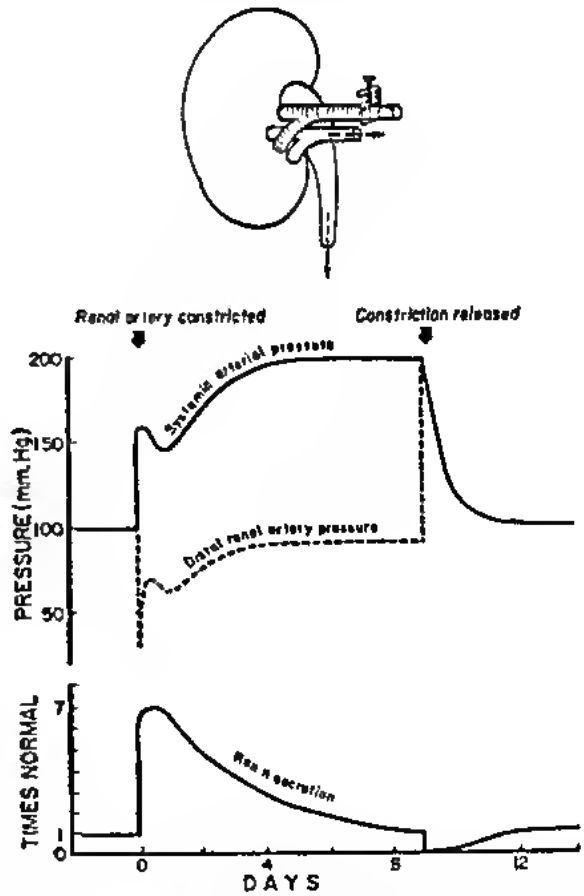


Figure 22–13. Effect of placing a constricting clamp on the renal artery of one kidney after the other kidney has been removed. Note the changes in systemic arterial pressure, renal artery pressure distal to the clamp, and rate of renin secretion. The resulting hypertension is called "one kidney Goldblatt hypertension."

sodium and water. Thus, both kidneys, but for different reasons, become sodium and water retainers. Consequently, hypertension develops. Furthermore, the excessive secretion of renin by the ischemic kidney never ceases, because the "normal" kidney will not retain quite enough salt and water to allow the arterial pressure to rise high enough ever to return the blood flow to the ischemic kidney entirely to normal.

Hypertension Caused by Diseased Kidneys That Secrete Renin Chronically. Often patchy areas of one or both kidneys are diseased and ischemic while other areas of the kidneys are normal. When this occurs, the almost identical effects occur as in the two-kidney type of Goldblatt hypertension. That is, the patchy ischemic kidney tissue secretes renin, and this in turn, acting through the formation of angiotensin II, causes the remaining kidney mass also to retain salt and water. Indeed, one of the most common causes of renal hypertension is such patchy ischemic kidney disease.

Hypertension In Toxemia of Pregnancy

During pregnancy many patients develop hypertension, which is one of the manifestations of the syndrome called *toxemia of pregnancy*. The pathological abnormality that causes this hypertension is believed to be thickening of the glomerular membranes, which reduces the rate of fluid filtration from the glomeruli into the renal tubules. For obvious reasons, the pressure level of the renal output curve is elevated, and the long-term level of arterial pressure becomes correspondingly elevated. These patients are especially prone to hypertension when they eat large quantities of salt.

Neurogenic Hypertension

Acute hypertension can be caused by strong *stimulation of the sympathetic nervous system*. For instance, when a person becomes excited for any reason or at times during states of anxiety, the sympathetic system becomes excessively stimulated and acute hypertension ensues.

Acute Neurogenic Hypertension Caused by Sectioning the Baroreceptor Nerves. Another type of acute neurogenic hypertension occurs when the nerves leading from the baroreceptors are cut or when the tractus solitarius is destroyed in each side of the medulla oblongata (these are the areas where the nerves from the baroreceptors connect in the brain stem). The sudden cessation of normal nerve signals from the baroreceptors has the same effect on the nervous pressure control mechanisms as a sudden reduction of the arterial pressure in the aorta and in the carotid arteries. That is, the vasomotor center suddenly becomes extremely active and the mean arterial pressure rises from 100 mm Hg to as high as 160 mm Hg. However, the pressure returns to normal within two to three days because the response of the vasomotor center to the absent baroreceptor signals fades away, which is called central "resetting" of the baroreceptor pressure control mechanism. Therefore, the neurogenic hypertension caused by sectioning the baroreceptor nerves is only an acute type of hypertension, not a chronic type.

Possible Neurogenic Mechanisms in Chronic Hypertension. Despite the well-documented specific instances of *acute* hypertension caused by neurogenic mechanisms, it has been difficult to prove that abnormalities of the nervous system can cause chronic hypertension. Nevertheless, most clinicians have the "clinical impression" that prolonged nervous tension can maintain prolonged increase in sympathetic stimulation of the blood vessels and kidneys and therefore can lead to chronic elevation of the arterial pressure.

However, probably more likely as a neurogenic mechanism of chronic hypertension is the following: Frequently, the sympathetic nervous system becomes extremely activated for short periods of time. During such periods, renal blood flow is often decreased to extremely low levels because of severe renal vascular vasoconstriction. Recent experiments have demonstrated that infusing norepinephrine, the sympathetic transmitter substance, directly into the renal arteries can cause permanent damage to the kidneys that can possibly lead to a prolonged renal type of hypertension. Thus, it is possible that episodes of extreme sympathetic activity, which are likely to occur in anyone's life, could under appropriate conditions lead to renal organic changes that could be the basis for permanent hypertension.

Spontaneous Hypertension in Lower Animals

Spontaneous, hereditary hypertension has been observed in a number of different strains of lower animals, including four different strains of rats, at least one strain of rabbits, and at least one strain of dogs. In the strain of rats that has been studied to the greatest extent, the *Okamoto strain*, there is evidence that in the early development of hypertension the sympathetic nervous system is considerably more active than in normal rats. However, in the late stages of this type of hypertension, structural changes have been observed in the nephrons of the kidneys. Therefore, these structural changes could easily be the basis for the long-term continuance of the hypertension. In two other strains of hypertensive rats, changes in renal function also have been observed.

ESSENTIAL HYPERTENSION

Approximately 90 per cent of all persons who have hypertension are said to have "essential hypertension." This term means simply that *the hypertension is of unknown origin*. However, in most patients with essential hypertension, there is a very strong hereditary tendency, the same as occurs in strains of hypertensive lower animals.

Some of the characteristics of severe essential hypertension are the following:

(1) The mean arterial pressure is increased about 40 to 50 per cent.

(2) The renal blood flow in the later stages is decreased to about one-half normal.

(3) The resistance to blood flow through the kidneys is increased three- to four-fold.

(4) Despite the great decrease in renal blood flow, the glomerular filtration rate is often very near normal. The reason for this is that the high arterial pressure still causes adequate filtration of fluid through the glomeruli into the renal tubules.

(5) The cardiac output is approximately normal.

(6) The total peripheral resistance is increased about

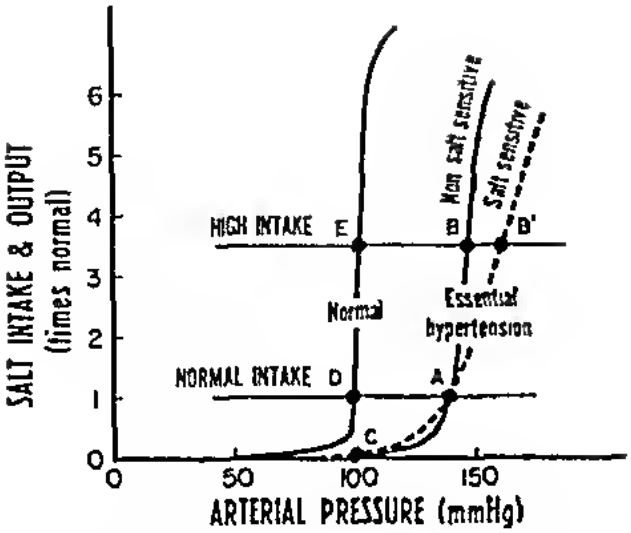


Figure 22–14. Analysis of arterial pressure regulation in "non salt sensitive" essential hypertension and "salt sensitive" essential hypertension. Explained in the text. (From Guyton et al. *Annu. Rev. Med.* 31:15, 1980.)

40 to 50 per cent, about the same amount that the arterial pressure is increased.

And, finally, the most important finding of all in persons with essential hypertension is the following:

(7) The kidneys will not excrete adequate amounts of salt and water unless the arterial pressure is high. In other words, if the mean arterial pressure in the essential hypertensive person is 150 mm Hg, reduction of the arterial pressure artificially to the normal value of 100 mm Hg (but without otherwise altering renal function except for the decreased pressure) will cause almost total anuria, and the person will retain salt and water until the pressure rises back to the elevated value of 150 mm Hg.

The exact reason for this failure of the kidneys of essential hypertensive persons to excrete salt and water at normal pressure levels is unknown. However, because of very significant vascular changes in the kidneys, as previously noted, one can surmise that decreased renal blood flow is the cause of this.

Graphical Analysis of Arterial Pressure Control in Essential Hypertension. Figure 22–14 is a graphical analysis of essential hypertension. The curves of this figure are not the same renal output curves that have been used in the early figures of this chapter. Instead, these curves are the sodium-pressure relationship that was shown for the normal kidneys by the heavy curve in Figure 22–9. The reason for using this relationship rather than the renal output curve is the following: It is very easy to determine the sodium-pressure relationship by simply increasing the level of sodium intake to a new level every few days and recording the changes in arterial pressure. When this procedure is employed in essential hypertensive patients, the two curves to the right in Figure 22–14 are recorded in two different groups of essential hypertensive patients, one called (1) *nonsalt-sensitive* patients, and the other (2) *salt-sensitive*. Note in both instances that the sodium-pressure relationship is shifted to a much higher pressure level than this relationship for normal persons. Now, let us plot on this same graph (a) a normal level of salt intake, and (b) a high level of salt intake representing 3.5 times the normal intake. In the case of the person with nonsalt-sensitive essential hypertension, the arterial pressure does not increase significantly when changing from normal salt intake to high salt intake. On the other hand, in those patients who have salt-sensitive essential hypertension, the high salt intake further exacerbates the hypertension. The reason for the difference between nonsalt-sensitive essential hypertension and salt-sensitive hypertension is presumably structural or functional differences in the kidneys of these two types of hypertensive patients.

Treatment of Essential Hypertension. Essential hypertension generally can be treated by two different types of drugs: (1) drugs that will increase renal blood flow, or (2) drugs that decrease tubular reabsorption of salt and water.

Those drugs that will increase renal blood flow are the various *vasodilator drugs*. Different ones act in one of the following three ways: (a) by inhibiting sympathetic nervous signals to the kidneys or blocking the action of the sympathetic transmitter substance on the renal vasculature, (b) by directly paralyzing the smooth muscle of the renal vasculature, or (c) by blocking the action of the renin-angiotensin system on the renal vasculature.

Those drugs that reduce reabsorption of salt and water by the tubules include especially drugs that block active transport of sodium through the tubular wall; this blockage in turn also prevents the reabsorption of water as well, as was explained earlier in the chapter. Such drugs are called *natriuretics* or *diuretics*; they will be discussed in greater detail in Chapter 38.

EFFECTS OF HYPERTENSION ON THE BODY

Hypertension can be very damaging because of two primary effects: (1) increased work load on the heart and (2) damage to the arteries themselves by the excessive pressure.

Effects of Increased Work Load on the Heart. Cardiac muscle, like skeletal muscle, hypertrophies when its work load increases. In hypertension, the very high pressure against which the left ventricle must beat causes it to increase in weight as much as two- to threefold. This increase is not accompanied by quite as much increase in coronary blood supply as there is increase in muscle tissue itself. Therefore, *relative ischemia* of the left ventricle develops as the hypertension becomes more and more severe. In the late stages of hypertension, this can become serious enough that the person develops angina pectoris. Also, the very high pressure in the coronary arteries causes rapid development of coronary arteriosclerosis so that hypertensive patients tend to die of coronary occlusion at much earlier ages than do normal persons.

Effects of the High Pressure in the Arteries. High pressure in the arteries not only causes coronary sclerosis but also sclerosis of blood vessels throughout all the remainder of the body. The arteriosclerosis process causes blood clots to develop in the vessels and also causes the blood vessels to become weakened. Therefore these vessels frequently thrombose, or they rupture and bleed severely. In either case, marked damage can occur in organs throughout the body. The two most important types of damage that occur in hypertension are:

(1) *Cerebral hemorrhage*, which means bleeding of a cerebral vessel with resultant destruction of local areas of brain tissue, and

(2) *Hemorrhage of renal vessels* inside the kidneys, which destroys large areas of the kidneys and therefore

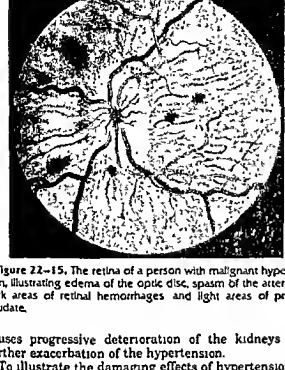


Figure 22–15. The retina of a person with malignant hypertension, illustrating edema of the optic disc, spasm of the arterioles, dark areas of retinal hemorrhages and light areas of protein exudate.

causes progressive deterioration of the kidneys and further exacerbation of the hypertension.

To illustrate the damaging effects of hypertension on the small vessels of the microcurculation, Figure 22–15 shows the retina of a person with very severe hypertension called *malignant hypertension*. This figure shows edema of the optic disc, spasm of the arterioles, dark areas in the retina that are the result of hemorrhages from the small arteries, and very light areas in the retina that represent protein that has exuded out of the blood vessels into the surrounding tissues.

REFERENCES

Anderson, D E.. Interactions of stress, salt, and blood pressure *Annu Rev. Physiol*, *46* 143, 1984

Antonaccio, M J Angiotensin converting enzyme (ACE) inhibitors *Annu Rev Pharmacol. Toxicol*, *22* 57, 1982

Bernstein, K N, and D'Connor, D T Antiadrenergic antihypertensive drugs. Their effect on renal function *Annu. Rev Pharmacol Toxicol*, *24* 105, 1984

Bianchi, G., *et al*.. Sodium balance and peripheral resistance in arterial hypertension *J Cardiovasc Pharmacol.* 6(Suppl 2):S457, 1984

Blaustein, M. P, and Hamlyn, J M. Sodium transport inhibition, cell calcium, and hypertension The natriuretic hormone/Na⁺–Ca⁺ ex change/hypertension hypothesis *Am J Med.*, *77* 45 1984

Bergstrom, P, *et al*., An evaluation of the metabolic interaction with myogenic vascular reactivity during blood flow autoregulation *Acta Physiol Scand*, *122* 275, 1984

Brody, M. J, *et al*, Neural mechanisms in hypertension *Annu Rev Physiol*, *42* 441, 1980

Canessa, M The polymorphism of red cell Na and K transport in essential hypertension Findings, controversies, and perspectives. *Prog Clin. Biol Res*, *159* 293, 1984

Cauvin, C, *et al*.. Mechanisms of calcium antagonist-induced vasodilation *Annu Rev Pharmacol Toxicol*, *23* 373, 1983

Coleman, T. G, and Guyton, A C Hypertension caused by salt loading in the dog III. Onset transients of cardiac output and other circulatory variables. *Circ Res*, *25* 153, 1969

Coleman, T G., and Guyton, A C The pressor role of angiotensin in salt deprivation and renal hypertension. *Clin Sci*, *48* 455–488, 1975

Coleman, T. G, *et al*.. Regulation of arterial pressure in the anephric state *Circulation*, *42* 569, 1970

Coleman, T G., *et al*: Experimental hypertension and the long term control of arterial pressure *In* MTP International Review of Science Physiology Baltimore, University Park Press, 1974, Vol 1, p 259.

Conway, J Hemodynamic aspects of essential hypertension in humans. *Physiol Rev*, *64* 617, 1984

Cowley, A. W, Jr, *et al* Open-loop analysis of the renin angiotensin system in the dog *Circ. Res*, *28* 568, 1971.

DeClue, J W, *et al*.. Subpressor angiotensin infusion, renal sodium handling, and salt-induced hypertension in the dog *Circ Res*, *43*(4) 503, 1978

Folkow, B. *Physiological aspects of primary hypertension Physiol Rev*, *62* 347 1982

Frais, E D Hemodynamics of hypertension. *Physiol Rev.*, *40*:27, 1960

Gibbons G H, *et al*: Interaction of signals influencing renin release *Annu Rev Physiol.*, *46* 231, 1984

Granger, H J, and Guyton, A. C.. Autoregulation of the total systemic circulation following destruction of the central nervous system in the dog. *Circ Res*, *25*:379, 1969

Guyton, A C Acute hypertension in dogs with cerebral ischemia. *Am. J Physiol*, *154* 45, 1948

Guyton, A C Essential cardiovascular regulation—the control linkages between bodily needs and circulatory function. *In* Dickinson, C J, and Marks, J (eds) Developments in Cardiovascular Medicine. Lancaster, England, MTP Press, 1978, p 265

Guyton, A C Arterial Pressure and Hypertension. Philadelphia, W. B Saunders Co, 1980

Guyton, A. C, and Coleman, T G Long-term regulation of the circulation. Interrelationships with body fluid volumes. *In* Physical Bases of Circulatory Transport Regulation and Exchange Philadelphia, W B Saunders Co, 1967

Guyton, A C, and Coleman, T G Quantitative analysis of the pathophysiology of hypertension *Circ Res*, *24* 1, 1969

Guyton, A C, *et al* Physiological control of arterial pressure *Bull N Y Acad Med*, *45* 811, 1969

Guyton, A C, *et al*. Arterial pressure regulation. Overriding dominance of the kidneys in long term regulation and in hypertension *Am. J Med* *52* 584, 1972

Guyton, A C, *et al* A systems analysis approach to understanding long range arterial blood pressure control and hypertension *Circ Res*, *35* 159, 1974

Guyton, A. C, *et al*., Integration and control of circulatory function. *Int. Rev Physiol*, *9*:341, 1976

Guyton, A C, *et al*. Salt balance and long-term blood pressure control *Annu Rev Med*, *31* 15–27, 1980

Hadley, F J The role of a humoral Na⁺, K⁺–ATPase inhibitor in regulating precapillary vessel tone *J Cardiovasc Pharmacol*, 6(Suppl 2) S439, 1984

Hall J E, *et al* Renal hemodynamics in acute and chronic angiotensin II hypertension *Am J Physiol.*, *235*(3):F174, 1978

Hall, J E, *et al* Control of arterial pressure and renal function during glucocorticoid excess in dogs *Hypertension*, *2* 139, 1980

Insel, P A, and Snavely, M D Catecholamines and the kidney Receptors and renal function *Annu Rev Physiol*, *43* 625, 1981

Kaplan, N M Non-drug treatment of hypertension *Annu Intern Med.*, *102*:359, 1985

Langston, J B, *et al* Effect of changes in salt intake on arterial pressure and renal function in partially nephrectomized dogs. *Circ Res*, *12* 508, 1963

Laird, J F, *et al* Renin aldosterone, body fluid volumes, and baroreceptor reflex in the development and reversal of Goldblatt hypertension in conscious dogs *Circ Res*, *34* 549, 1974

Lohmeier, T E, *et al*: Failure of chronic aldosterone infusion to increase arterial pressure in dogs with angiotensin induced hypertension. *Circ Res*, *43*(3):381, 1978

Manning R. D, Jr, *et al* Essential role of mean circulatory filling pressure in salt-induced hypertension *Am J Physiol*, *236* R40, 1979

McAllister, R G, Jr., *et al* Pharmacokinetics of calcium-entry blockers. *Am J Cardiol*, *55* 39B, 1985

Norman, R A, Jr, *et al* Arterial pressure–urinary output relationship in hypertensive rats. *Am. J Physiol*, *234*(3) R98, 1978

Norman, R. A., Jr, *et al*.. Renal function curves in normotensive and spontaneously hypertensive rats. *Am. J. Physiol.*, *234* R98, 1978

Reed, G, and Devous, M. Cerebral blood flow autoregulation and hypertension *Am J. Med. Sci*., *289*:37, 1985.

Robinson, B F.. Calcium-entry blocking agents in the treatment of systemic hypertension *Am. J Cardiol*, *55* 102B, 1985

Sant, S L: Pulmonary metabolism of prostaglandins and vasoactive peptides *Annu Rev Physiol*, *44* 257, 1982

Schmid, P G, *et al* Peptides and blood vessels *In* Shepherd, J. T., and Abboud, F M (eds). Handbook of Physiology Sec. 2, Vol III. Bethesda, American Physiological Society, 1983, p 815

Sterns, R. H., *et al* 'Page kidney' Hypertension caused by chronic subcapsular hematoma. *Arch Intern Med.*, *145* 169, 1985

Sullivan, J. M.. Drug suppression of the angiotensin system in congestive heart failure *Annu Rev. Med.*, *34* 169, 1983

Torretta, J., Sympathetic control of renin release *Annu. Rev. Pharmacol. Toxicol*, *22* 167, 1982.

Valtre, B L., *et al*.. Tumor-derived angiogenesis factors from rat Walker 256 carcinoma. An experimental investigation and review *Experientia*, *41* 1, 1985

Vanhoutte, P. M: Calcium-entry blockers, vascular smooth muscle and systemic hypertension *Am J. Cardiol*, *55* 17B, 1985

Young, D B, and Guyton, A. C.: Steady state aldosterone dose response relationships. *Circ Res*., *40*(2) 138, 1977

conditions as myocardial infarction, valvular heart disease, myocarditis, and congenital heart abnormalities can all reduce the pumping effectiveness of the heart. In these instances, the permissive level for heart pumping may fall below 5 liters per minute, indeed for a few hours even as low as 2 to 3 liters per minute. When this happens, the heart becomes unable to cope with the amount of blood that is attempting to flow into the right atrium from the peripheral circulation. Therefore, the heart is said to *fail*, meaning simply that it fails to pump the amount of blood that is demanded of it. Under these conditions, the heart becomes the limiting factor in cardiac output control. However, this is not the normal state. We shall discuss cardiac failure in Chapter 26.

ROLE OF TOTAL PERIPHERAL RESISTANCE IN DETERMINING NORMAL VENOUS RETURN AND CARDIAC OUTPUT

Let us recall Ohm's law for calculating blood flow through blood vessels that was presented in Chapter 18:

$$\text{Blood flow} = \frac{\text{Input pressure} - \text{Output pressure}}{\text{Resistance}}$$

Now if we apply this to blood flow through the systemic circulation, we can determine venous return

$$\text{Venous return} = \frac{\text{Arterial pressure} - \text{Right atrial pressure}}{\text{Total peripheral resistance}}$$

And since the right atrial pressure remains very nearly zero, this formula becomes.

$$\text{Venous return} = \frac{\text{Arterial pressure}}{\text{Total peripheral resistance}}$$

Therefore, the amount of blood that flows into the heart each minute (venous return) and that is pumped each minute (cardiac output) is determined by two prime factors (1) the arterial pressure, and (2) the total peripheral resistance. When the arterial pressure remains normal, as it usually does, venous return and cardiac output are then inversely proportional to the total peripheral resistance. To state this still another way, every time a peripheral blood vessel dilates, the venous return and cardiac output increase. Furthermore, the more vessels that dilate, the greater the cardiac output becomes.

Cardiac Output Regulation as the Sum of the Local Blood Flow Regulations Throughout the Body. As long as the arterial pressure remains normal, each local tissue in the body can control its own blood flow by simply dilating or constricting its local blood vessels. This mechanism, which was discussed in detail in Chapter 20, is the means by which each tissue protects its own nutrient supply, controlling the blood flow in response to its own needs.

Therefore, since the venous return to the heart is the sum of all the local blood flows through all the individual tissues of the body, the local blood flow regulatory mechanisms in the peripheral circulation are the true controllers of cardiac output under normal conditions. This is an automatic mechanism that allows the heart to respond instantaneously to the needs of each individual tissue. If some tissues need extra blood flow and their local blood vessels dilate, the venous return increases automatically, and the cardiac output increases by an equivalent amount. If all the tissues throughout the body require increased blood flow at the same time, the venous return becomes very great, and the cardiac output increases accordingly.

Thus, the whole theory of normal cardiac output regulation is that *the tissues control the output* in accordance with their needs. Again, it must be stated that it is not the heart that controls the cardiac output under normal conditions; instead, the heart plays a permissive role that allows the tissues to do the controlling. The heart does this by maintaining a permissive pumping capacity that is somewhat above the actual venous return—that is, except when the heart fails.

Effect of Local Tissue Metabolism on Cardiac Output Regulation, and Importance of Simultaneous Arterial Pressure Regulation. The most important factor that controls the local blood flows in the individual tissues is the metabolic rates of the respective tissues. Therefore, venous return and cardiac output are normally controlled in relation to the level of metabolism of the body. This effect was illustrated in Figure 23–1, which showed that the cardiac output increases directly in relation to work output during exercise and that the cardiac output increases parallel to the increase in oxygen consumption, which is a measure of the rate of metabolism.

However, it is essential that the arterial pressure be maintained at a normal level if changes in metabolism are to regulate cardiac output. This is illustrated in Figure 23–3, which shows changes in cardiac output in response to an approximately five-fold increase in tissue metabolism in a dog. The increase in metabolism was caused by the toxic substance dinitrophenol, which causes the metabolic systems to increase their metabolic use of oxygen tremendously and, simultaneously, to dilate the local blood vessels supplying the tissues. Note in this figure that when the arterial pressure was normally controlled, the cardiac output increased approximately 300 per cent. On the other hand, when the nervous mechanisms for control of arterial pressure were blocked, the cardiac output increased less than 100 per cent.

Therefore, once again, we can state that, under

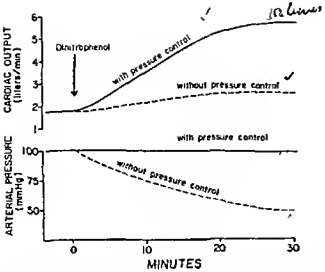


Figure 23–3. An experiment in a dog to demonstrate the importance of arterial pressure control as a prerequisite for cardiac output control. Note that with pressure control the metabolic stimulant dinitrophenol increases cardiac output; without pressure control, the arterial pressure falls and the cardiac output rises very little. (Drawn from experiments by Dr M. Banet.)

normal conditions, venous return and cardiac output are determined almost entirely by the degree of dilatation of the local blood vessels in the tissues throughout the body. However, this mechanism operates properly only in case the mechanisms for maintaining a normal arterial pressure, which were discussed in the preceding two chapters, also function properly.

EFFECT OF ARTERIOVENOUS FISTULAE ON CARDIAC OUTPUT

The effect of an A-V fistula on cardiac output is discussed here because of the extremely important lessons that can be learned from this condition. An A-V fistula is a direct opening from an artery into a vein that allows rapid flow of blood directly from the artery to the vein. In experimental animals, such a fistula can be opened and closed artificially, and one can study the changes in fistula flow, cardiac output, and arterial pressure. A typical experiment of this type in the dog is illustrated in Figure 23–4, showing an initial cardiac output of 1300 ml per minute. After 15 seconds of recording normal conditions, a fistula from the aorta to the vena cava is opened through which 1200 ml of blood flows per minute. Note that the cardiac output increases instantaneously about 1000 ml per minute—about 84 per cent as much increase in cardiac output as there is fistula flow. (If the fistula is left open for 24 hours or more, the cardiac output increases 100 per cent as much as the fistula flow because of renal retention of fluid that returns the initially decreased arterial pressure all the way to normal.) Then, when the fistula is closed, the cardiac output returns back to its original level.

Now for the lesson to be learned from this study: When the fistula is opened, the pumping ability of the heart is not changed at all; there is only a change in the peripheral circulation—opening of the A-V fistula. Nevertheless, within a few heartbeats after the fistula is opened, the cardiac output increases almost as much as the increase in blood flow through the fistula. In other words, the fistula flow is added to the flow through the remainder of the body, and the two of these summate to determine the venous return and the cardiac output.

This experiment demonstrates the two cardinal features of normal cardiac output regulation: (1) almost all normal cardiac output regulation is determined by changes in total peripheral resistance, and (2) the normal heart has a permissive level of pumping that is considerably above the actual cardiac output. The heart, therefore, does not have to be stimulated for the increase in cardiac output to come about unless more cardiac output is required than the heart can pump without stimulation.

SOME SPECIAL PROBLEMS IN CARDIAC OUTPUT REGULATION

Thus far wa have discussed cardiac output regulation when all aspects of circulatory function are normal and when the circulatory system is not stressed beyond moderate limits. Also, we hava assumed that the arterial pressure-regulating mechanisms continue to maintain the arterial pressure at a normal level, that the blood voluma is normal, and that the heart is normal.

On the other hand, when the cardiac output approaches the limit that the heart can pump, as occurs in very strenuous exercise, or when blood volume or other circulatory factors become abnormal, then it becomes necessary to consider other

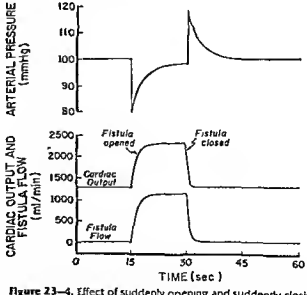


Figure 23–4. Effect of suddenly opening and suddenly closing an A-V fistula, showing changes in fistula flow, cardiac output, and arterial pressure.

important effects in cardiac output regulation, as follows:

ROLE OF BLOOD VOLUME AND THE "MEAN SYSTEMIC FILLING PRESSURE" IN CARDIAC OUTPUT REGULATION

If the quantity of blood in the circulatory system is too little to fill the system adequately, blood will flow very poorly from the peripheral vessels into the heart. Therefore, the degree of filling of the circulation then becomes a critical factor in determining the venous return to the heart and therefore also in determining the cardiac output.

The student may recall from Chapter 18 that the degree of filling of the circulatory system with blood can be expressed in terms of the *mean circulatory filling pressure*, and that the degree of filling of the systemic portion of the circulation can be expressed in terms of the *mean systemic filling pressure*. The mean systemic filling pressure is the pressure measured in the systemic circulation when blood flow in the circulatory system is suddenly stopped and the blood in the systemic vessels is redistributed so that the pressure in all the vessels is equal. Normally, the mean systemic filling pressure is 7 mm Hg, but an increase in blood volume of 15 to 30 per cent doubles the mean systemic filling pressure, whereas a decrease in blood volume of this same amount reduces the mean systemic filling pressure to zero.

The mean systemic filling pressure is the *average* effective pressure of the blood in the peripheral circulation that tends to push the blood toward the heart. Experiments have demonstrated that, *when the peripheral resistance remains constant*, the rate of *venous return* from the systemic vessels through the veins to the heart is directly proportional to the *mean systemic filling pressure* minus *the right atrial pressure*. Therefore, whenever the mean systemic filling pressure falls to zero, the flow of blood returning to the heart likewise approaches zero.

Consequently, it is very important that the mean systemic filling pressure remain high enough at all times to supply the peripheral pressure needed to push blood from the peripheral vessels back to the heart. The quantitative relationship of the mean systemic filling pressure to cardiac output regulation will be considered in much more detail later in the chapter in connection with graphical analysis of cardiac output regulation.

REGULATION OF CARDIAC OUTPUT IN HEAVY EXERCISE, WHICH REQUIRES SIMULTANEOUS PERIPHERAL AND CARDIAC ADJUSTMENTS

Heavy exercise is one of the most stressful conditions to which the body is ever subjected. The tissues can require as much as 20 times normal [illegible] of oxygen and other nutrients, so that simply to transport enough oxygen from the lungs to the tissues sometimes demands a minimum cardiac output increase of five- to sixfold. This is greater than the amount of cardiac output the normal, unstimulated heart can pump. Therefore, to insure the massive increase in cardiac output that is required during heavy exercise, almost all factors that are known to increase cardiac output must be called into play.

Muscle Vasodilatation Resulting from Increased Metabolism. By far the most important factor that increases the cardiac output during exercise is the vasodilatation that occurs in the exercising muscles. The cause of the vasodilatation is the large increase in muscle metabolism during exercise. This leads to tremendously increased use of oxygen and other nutrients by the muscles, as well as formation of vasodilator substances, all of which act together to cause marked local vascular dilatation and greatly increased local blood flow. This local vasodilatation requires 5 to 15 seconds to reach full development after a person begins to exercise strongly; but once it does reach full development, the very great decrease in vascular resistance allows extreme quantities of blood to flow into the veins and thence to the heart, thereby markedly increasing venous return and cardiac output.

Role of the Heart in Strenuous Exercise. When large numbers of muscles are exercising simultaneously, the peripheral vasodilatation may be so great and the venous return to the heart so voluminous that the "resting" heart simply cannot pump this extra amount of blood. Therefore, it is essential that the *permissive level of pumping* by the heart be greatly increased from its normal level of 10 to 13 liters per minute. This is achieved mainly by sympathetic stimulation of the heart (but also partly by a *decrease* in parasympathetic stimulation), which increases the permissive level to as high as 20 to 25 liters per minute in a normal person or as high as 35 liters in the trained athlete.

Therefore, in exercise, the cardiac output can increase two- to threefold without stimulation of the heart. But to go above this amount it is absolutely essential that the heart be nervously excited to increase the permissive level of heart pumping.

Special Functions of the Sympathetic Nervous System During Exercise. In light exercise, the heart and the circulation do not need to be stimulated by the sympathetic nervous system. But in heavy exercise, sympathetic stimulation is an absolute essential. Sympathetic stimulation has multiple effects on the circulation that are critical in increasing the cardiac output to the very high levels required in heavy exercise. These effects are:

1. Strong sympathetic stimulation can increase the permissive level of heart pumping from 10 to 13 liters per minute to as high as 20 to 25 liters

per minute (or even higher in the trained athlete), as already noted.

2. Sympathetic stimulation constricts almost all the blood vessels throughout the body besides those in the exercising muscles. The veins, especially, are constricted, which increases the mean systemic filling pressure to as much as 2.5 times normal This increase in mean systemic filling pressure provides an extra push on the blood in the peripheral vessels to return this blood to the heart. This is absolutely essential to achieve enough venous return, and therefore to achieve enough cardiac output, in heavy exercise.

3. It is believed by some physiologists that a special component of the sympathetic nervous system, a so-called "vasodilator" component, is stimulated during exercise to cause direct vasodilatation of the muscle blood vessels. This, theoretically, could further enhance blood flow through the muscles and thereby help supply the required nutrients, but its effects are still in doubt

Mechanisms for Stimulating the Sympathetic Nervous System During Exercise. It is believed that at least three different mechanisms enhance sympathetic activity during exercise.

1. Simply thinking about exercise has the psychic effect of exciting the autonomic nervous centers, which increases the heart rate, increases the strength of heart contraction, and constricts the blood vessels throughout the body to increase the mean systemic filling pressure. Together, these effects can increase the cardiac output instantaneously as much as 50 per cent even before the exercise begins.

2. The signals from the motor cortex that excite activity in the muscles are believed also to send collateral signals into the vasomotor center to excite the sympathetic nervous system. This causes vasoconstriction in most tissues of the body besides in the exercising muscles, thus increasing the arterial pressure. The increase in pressure in turn provides an extra force to push blood through the muscles. Also, the signals from the motor cortex are believed to stimulate *sympathetic vasodilator fibers* that go directly to the muscles and increase muscle blood flow. It seems certain that this vasodilator mechanism operates during the first few seconds of exercise to provide increased blood flow before the local metabolic vasodilatation occurs in the muscles, but it is probably much less important once the *local vasodilatation* has occurred.

3. Sensory nerve endings in the muscles are believed to be excited by metabolic products produced during muscle contraction, and signals then pass to the vasomotor center to excite the sympathetic system still more. In fact, some physiologists believe that this stimulation of the sympathetic nervous system is the most important of the sympathetic stimulatory mechanisms, but this is not yet certain.

Role of Abdominal Contraction for Enhancing Venous Return and Cardiac Output. At the very beginning of exercise, the initial tenseness of the body that is associated with exercise causes abdominal contraction and compression of the large venous reservoirs in the abdomen. This immediately increases the mean systemic filling pressure two- to threefold, which can increase venous return and cardiac output as much as 30 to 80 per cent within one to two beats of the heart. This thus provides an "anticipatory" increase in cardiac output even before the muscles require the extra blood flow.

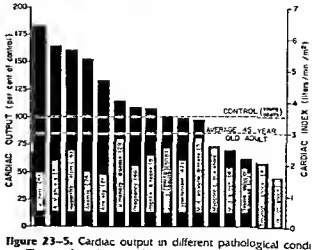
Summary. It is clear that a complex set of controls comes into play during exercise to allow the heart and the circulation to increase the cardiac output to the tremendous levels required for nutrient supply to the muscles The increase in cardiac output occurs first when the person begins to anticipate exercise, which stimulates the sympathetic nervous system. Then, tensing of the abdominal muscles gives an additional surge of venous return that also increases the cardiac output. Finally, within seconds after the exercise actually begins, intense local metabolic vasodilatation occurs in the muscles themselves. Combined with the hyperdynamic state of the heart and the increased mean systemic filling pressure, this causes the further surge in the cardiac output up to levels between 20 and 35 liters per minute.

ABNORMALLY LOW AND ABNORMALLY HIGH CARDIAC OUTPUTS

Low Output

Figure 23–5 illustrates the cardiac outputs in different pathological conditions, *showing at the far right the* conditions that cause abnormally low outputs and at the left the conditions that cause abnormally high outputs.

It is easy to understand those abnormalities of the


CARDIAC OUTPUT (per cent of control)
CARDIAC INDEX (liters/min/m²)
200
175
150
125
100
75
50
25
0
7
6
5
4
3
2
1
0
CONTROL (1000)
AVERAGE 65 YEARS OLD ADULT
Hemorrhage (1)
Hemorrhage (1)
Hemorrhage (1)

circulation that cause low cardiac output These fall into two different categories. (1) those abnormalities that cause the pumping effectiveness of the heart to fall too low, and (2) those that cause venous return to fall too low

Decreased Cardiac Output Caused by Cardiac Factors. Whenever the heart becomes severely damaged, from whatever cause, its permissive level of pumping may fall below that needed for adequate blood flow to the tissues Some examples of this include *severe myocardial infarction, severe valvular heart disease, myocarditis, cardiac tamponade*, and certain *cardiac metabolic derangements* The effect of several of these is illustrated on the right in Figure 23–5, showing the low cardiac outputs that result

When the cardiac output falls so low that the tissues throughout the body begin to suffer nutritional deficiency, the condition is called *cardiac shock* This will be discussed fully in Chapter 26 in relation to cardiac failure

Decrease in Cardiac Output Caused by Peripheral Factors—Decreased Venous Return. Any factor that interferes with venous return also can lead to decreased cardiac output Some of these are.

1 *Decreased blood volume* By far the most common peripheral factor that leads to decreased cardiac output is decreased blood volume, resulting most often from hemorrhage It is very clear why this decreases the cardiac output. loss of blood decreases the mean systemic filling pressure to such a low level that there simply is not enough blood in the peripheral vessels to create the pressure required to push blood back to the heart In other words, the mean systemic filling pressure falls to or near to zero, so that venous return also approaches zero

2 *Acute venous dilatation* On some occasions the peripheral veins become acutely vasodilated. This results most often when the sympathetic nervous system suddenly becomes inactive For instance, fainting often results from sudden loss of sympathetic nervous system activity, which causes the peripheral capacitative vessels, especially the veins, to dilate markedly. This decreases the mean systemic filling pressure because the blood volume can no longer create adequate pressure in *the now flaccid peripheral blood vessels As a result the* blood "pools" in the vessels and will not return to the heart

3 *Obstruction of the veins.* On rare occasions, the large veins leading into the heart become obstructed, in which case the blood in the peripheral vessels cannot flow back into the heart. Consequently the cardiac output falls markedly

Regardless of the cause of low cardiac output, whether it be a peripheral factor or a cardiac factor, if ever the cardiac output falls below that level required for adequate nutrition of the tissues, the person is said to suffer *circulatory shock*. This can be lethal within a few minutes to a few hours Circulatory shock is such an important clinical problem that it will be discussed in detail in Chapter 28

High Cardiac Output—Role of Reduced Total Peripheral Resistance

The left side of Figure 23–5 identifies conditions that commonly cause cardiac outputs higher than normal. One of the distinguishing features of these conditions is *they all result from chronically reduced total peripheral resistance* And none of them result from excessive excitation of the heart itself, which we will explain below. For the present, let us look at some of the peripheral factors that can increase the cardiac output to above normal·

(1) *Beriberi.* This is a disease caused by insufficient quantity of the vitamin *thiamine* in the diet. Lack of this vitamin causes diminished ability of the tissues to utilize cellular nutrients, which in turn causes marked peripheral vasodilatotion The total peripheral resistance decreases sometimes to as little as one-half normal. Consequently, the long-term level of cardiac output also often increases to as much as two times normal

(2) *Arteriovenous fistula (shunt).* Earlier in the chapter we pointed out that whenever a fistula (also called a shunt) occurs between a major artery and a major vein, tremendous amounts of blood will flow directly from the artery into the vein This, too, greatly decreases the total peripheral resistance, and likewise increases the venous return and cardiac output.

(3) *Hyperthyroidism* In hyperthyroidism, the metabolism of all the tissues of the body becomes greatly increased. Oxygen usage increases, and vasodilator products are released from the tissues Therefore, the total peripheral resistance decreases markedly, and the cardiac output often increases to as much as 40 to 80 per cent above normal

(4) *Anemia* In anemia, two peripheral effects greatly decrease the total peripheral resistance One of these is reduced viscosity of the blood, resulting from the decreased concentration of red blood cells The other is diminished delivery of oxygen to the tissues because of the decreased hemoglobin, which causes local vasodilatation As a consequence, the total peripheral resistance decreases greatly, and the cardiac output increases

And, any other factor that decreases the total peripheral resistance chronically will also increase the cardiac output

Failure of Increased Cardiac Pumping to Cause Prolonged Increase of the Cardiac Output. If the heart is suddenly stimulated excessively, the cardiac output often increases as much as 20 to 30 per cent. However, even this small increase is maintained no longer than a few minutes even though the heart continues to be strongly stimulated There are two reasons for this (1) *Excess blood flow through the tissues causes* automatic vasoconstriction of the blood vessels because of the autoregulation mechanism discussed in previous chapters, and this reduces the venous return and cardiac output back toward normal. (2) The slightly increased arterial pressure that results following acute cardiac stimulation raises the capillary pressure, and fluid filters out of the capillaries into the tissues thereby decreasing the blood volume and also decreasing the venous return back toward normal. Also, the increased pressure causes the kidneys to lose fluid volume as well until the arterial pressure and cardiac output return to normal

Thus, all the known conditions that cause *chronic* elevation of the cardiac output result from decreased total peripheral resistance and not increased cardiac activity.

GRAPHICAL ANALYSIS OF CARDIAC OUTPUT REGULATION

The discussions of cardiac output regulation thus far are adequate for an understanding of the factors that

control cardiac output in most simple conditions, and they also present a reasonably complete *qualitative* picture of cardiac output control. However, it is often important to obtain a much more quantitative understanding of cardiac output regulation. For instance, in cardiac failure and circulatory shock, the qualitative type of analysis is only partially satisfactory. On the other hand, a more quantitative type of analysis will allow one to understand many details of these critical clinical conditions that are impossible to comprehend through qualitative thinking alone.

To perform quantitative analyses of cardiac output regulation, it is necessary to distinguish separately the two primary factors that are concerned with cardiac output regulation: (1) the pumping ability of the heart, and (2) the peripheral factors that affect flow of blood into the heart. Then, these two factors, the cardiac and the peripheral, must be put together in a quantitative way to see how they interact with each other to determine cardiac output. This process frequently has been accomplished by using electronic computers, but it can also be done more simply by using a type of graphical analysis that one can carry in the mind and that can be used daily in solving some of the quantitative problems of cardiac output regulation. In the following sections we will discuss this type of analysis. To begin, we will first characterize the pumping ability of the heart by use of *cardiac output curves*, and then we characterize the peripheral factors that affect venous return by use of *venous return curves*.

CARDIAC OUTPUT CURVES

Use of cardiac function curves to depict the ability of the heart to pump blood was discussed in Chapter 13, and the effects of several different factors on these function curves were demonstrated in Figures 13–9 and 13–10. For the purpose of understanding cardiac output regulation, the type of cardiac function curve required is that illustrated in Figures 23–6 and 23–7, which depicts the *cardiac output at different right atrial pressures* and is called the *cardiac output curve*. The normal cardiac output curve is labeled in each figure, whereas the other curves represent cardiac output in special conditions.

The basic factors that determine the precise characteristics of the cardiac output curve are (1) the *effectiveness of the heart as a pump*, and (2) the *extracardiac pressure*.

Effectiveness of the Heart as a Pump. A heart that is capable of pumping greater quantities of blood than can the normal heart, such as a heart strongly stimulated by sympathetic impulses or a hypertrophied heart, is called a *hypereffective heart*. A weakened heart is called a *hypoeffective heart*. Figure 23–6 illustrates the effect of hypereffectivity or hypoeffectivity on the cardiac output curves, showing that the plateau of the curve rises greatly in the case of the hypereffective heart and falls equally as much in the case of the hypoeffective heart. Obviously, there are all degrees of hypereffectivity and hypoeffectivity.

Factors That Can Cause a Hypereffective Heart. Factors that can make the heart a better pump than normal include:

(1) *Sympathetic stimulation*, which can increase the cardiac output curve to approximately two times normal.

(2) *Hypertrophy of the heart*, which can increase the cardiac output curve to perhaps as much as two times normal.

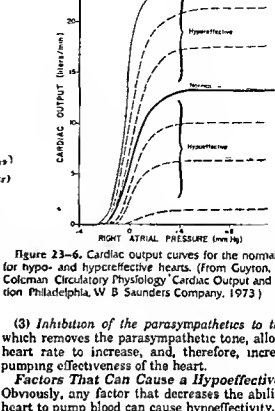


Figure 23–6. Cardiac output curves for the normal heart and for hypo- and hypereffective hearts. (From Guyton, Jones and Coleman Circulatory Physiology Cardiac Output and Its Regulation Philadelphia, W B Saunders Company, 1973.)

(3) *Inhibition of the parasympathetics to the heart*, which removes the parasympathetic tone, allowing the heart rate to increase, and, therefore, increases the pumping effectiveness of the heart.

Factors That Can Cause a Hypoeffective Heart. Obviously, any factor that decreases the ability of the heart to pump blood can cause hypoeffectivity. Some of the numerous factors are.

Myocardial infarction
Valvular heart disease
Increased arterial pressure against which the heart must pump
Parasympathetic stimulation of the heart
Inhibition of the sympathetics to the heart
Congenital heart disease
Myocarditis
Cardiac anoxia
Diphtheritic or other types of myocardial damage or toxicity

Effect of Extracardiac Pressure on Cardiac Output Curves. Figure 23–7 illustrates the effect of changes in extracardiac pressure on the cardiac output curve. The normal extracardiac pressure, which is also

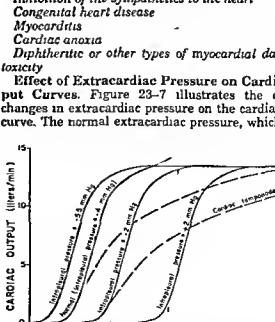


Figure 23–7. Cardiac output caused by changes in intrapleural pressure and by cardiac tamponade. (From Guyton, Jones, and Coleman. Circulatory Physiology. Cardiac Output and Its Regulation. Philadelphia, W. B. Saunders Company, 1973.)

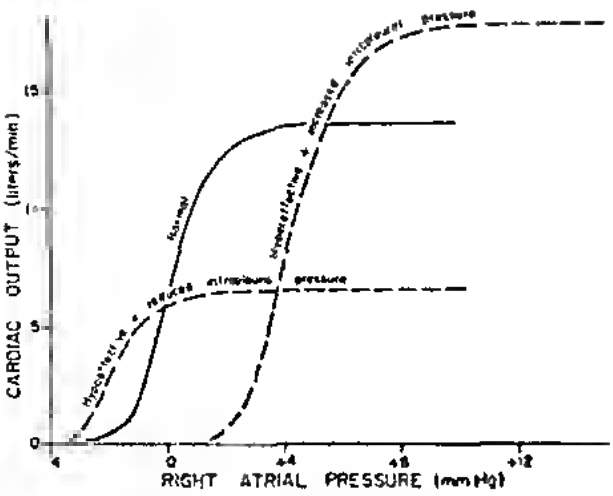


Figure 23–8. Combinations of two major patterns of cardiac output curves showing the effect of alterations in both extracardiac pressure and effectiveness of the heart as a pump (from Guyton, Jones, and Coleman: Circulatory Physiology: Cardiac Output and Its Regulation. Philadelphia, W. B. Saunders Company, 1973.)

equal to the normal intrapleural pressure, is −4 mm Hg. Note in the figure that a rise in intrapleural pressure to −2 mm Hg shifts the entire curve to the right by this same amount. This shift occurs because an extra 2 mm Hg right atrial pressure is now required to overcome the increased pressure on the outside of the heart. Likewise, an increase in intrapleural pressure to +2 mm Hg requires a 6 mm Hg increase in right atrial pressure, which obviously shifts the entire curve 6 mm Hg to the right. And, in the same manner, a more negative intrapleural pressure shifts the curve to the left.

Some of the factors that can alter the intrapleural pressure and thereby shift the cardiac output curve include:

(1) *Cyclic changes during normal respiration.*

(2) *Breathing against a negative pressure,* which shifts the curve to a more negative atrial pressure (to the left).

(3) *Positive pressure breathing,* which shifts the curve to the right.

(4) *Opening the thoracic cage,* which increases the intrapleural pressure and shifts the curve to the right.

(5) *Cardiac tamponade,* which means an accumulation of large quantities of fluid in the pericardial cavity around the heart with resultant increase in extracardiac pressure and shifting of the curve to the right. Note in Figure 23–7 that cardiac tamponade shifts the upper parts of the curves farther to the right than the lower parts because the extracardiac pressure rises to higher values as the chambers of the heart fill to increased volumes during high cardiac output.

Combinations of Different Patterns of Cardiac Output Curves. Figure 23–8 illustrates that the cardiac output curve can change as a result of simultaneous changes in extracardiac pressure and effectiveness of the heart as a pump. Thus, knowing what is happening to the extracardiac pressure and also the approximate capability of the heart as a pump, one can express the momentary ability of the heart to pump blood by a cardiac output curve.

VENOUS RETURN CURVES

There still remains the entire systemic circulation that must be considered before a total analysis of cardiac regulation can be made. To analyze the function of the systemic circulation, we remove the heart and lungs from the circulation of an animal and replace these with a pump and artificial oxygenator system. Then, different factors, such as changes in blood volume and changes in right atrial pressure, are altered to determine how the systemic circulation operates in different circulatory states. In these studies, one finds two principal factors that affect the function of the systemic circulation in relation to cardiac output regulation. These are (1) *the degree of filling of the systemic circulation,* which is measured by the *mean systemic filling pressure,* discussed in Chapter 18, and (2) the *resistance to blood flow* in the different segments of the systemic circulation.

The Normal Venous Return Curve. In the same way that the cardiac output curve relates cardiac output to right atrial pressure, the *venous return curve relates venous return also to right atrial pressure.*

The dark curve in Figure 23–9 is the normal venous return curve. This curve shows that as the right atrial pressure increases, it causes back pressure on the systemic circulation and thereby decreases venous return of blood to the heart. *If all circulatory reflexes are prevented from acting,* venous return decreases to zero when the right atrial pressure rises to equal the normal mean systemic filling pressure of about 7 mm Hg. Such a slight rise in right atrial pressure causes a drastic decrease in venous return because the systemic circulation is a very distensible bag, so that any increase in back pressure causes blood to dam up in this bag instead of returning to the heart. Lack of venous flow into the heart decreases the cardiac output, thereby decreasing the arterial pressure. Thus, at the same time that the right atrial pressure is rising, the arterial pressure is falling, and these two pressures come to equilibrium when all flow in the systemic circulation ceases at a pressure of 7 mm Hg, which, by definition, is the mean systemic filling pressure.

Plateau in the Venous Return Curve Caused by Collapse of the Veins. When the right atrial pressure falls *below* zero—that is, below atmospheric pressure—venous return does not increase significantly. By the

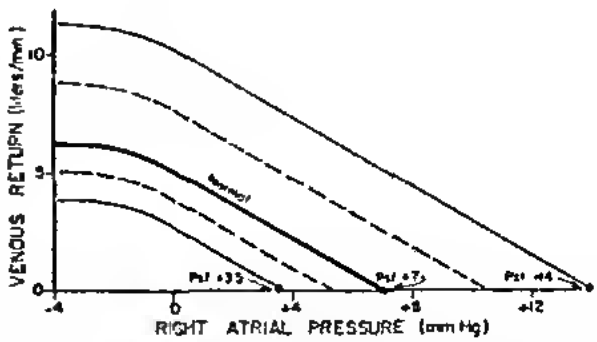


Figure 23–9. Venous return curves, showing the normal curve when the mean systemic filling pressure (*Psf*) is 7 mm Hg, and showing the effect of altering the mean systemic filling pressure (from Guyton, Jones, and Coleman: Circulatory Physiology: Cardiac Output and Its Regulation. Philadelphia, W. B. Saunders Company, 1973.)

time the right atrial pressure has fallen to about −2 mm Hg, the venous return will have reached a plateau; then it remains at this plateau level even though the right atrial pressure falls as low as −20 to −50 mm Hg. This plateau is caused by collapse of the veins entering the chest. Low right atrial pressure sucks the walls of the veins together where they enter the chest, which prevents the negative pressure from sucking blood through the veins. Instead, the pressure in the veins immediately outside the chest remains almost exactly equal to atmospheric pressure (zero pressure) Therefore, for all practical purposes, the peripheral venous pressure never falls below a value of 0 mm Hg despite the fact that the right atrial pressure may fall to very negative values.

Effect on the Venous Return Curve of Changes in Mean Systemic Filling Pressure. Figure 23–9 also illustrates the effect of increasing or decreasing the mean systemic filling pressure *(Psf)* By definition, the mean systemic filling pressure is the pressure in all parts of the systemic circulation when there is no flow in the circulation. Therefore, at the point where the venous return curve reaches the zero venous return level, the right atrial pressure becomes equal to the mean systemic filling pressure Note in Figure 23–9 that the normal mean systemic filling pressure is 7 mm Hg. For the uppermost curve in the figure, the mean systemic filling pressure is 14 mm Hg, and for the lowermost curve it is 3 5 mm Hg These curves demonstrate that the greater the mean systemic filling pressure, which also means the greater the "tightness" with which the circulatory system is filled with blood, the more the venous return curve shifts upward and to the right. Conversely, the lower the mean systemic filling pressure, the more the curve shifts downward and to the left. To express this another way, the greater the system is filled, the easier it is for blood to flow into the heart. And the less the filling, the more difficult it is for blooc to flow into the heart.

Factors That Can Alter the Mean Systemic Filling Pressure. The factors that can alter the mean systemic filling pressure were discussed in Chapter 18 The three primary factors are. (1) The mean systemic filling pressure increases rapidly with an *increase in blood volume—an acute increase in blood volume of 15 per cent or a chronic increase of about 30 per cent increases the mean systemic filling pressure to approximately double normal* (2) *Maximal sympathetic stimulation* can increase the mean systemic filling pressure from 7 mm Hg to about 17 mm Hg, approximately a 2 5 fold increase. (3) An *increase in contraction* of all the skeletal muscles throughout the body can increase the mean systemic filling pressure by compressing the blood vessels from the outside to about 25 mm Hg

Factors that decrease the mean systemic filling pressure include principally (1) *loss of sympathetic tone,* which decreases the mean systemic filling pressure from 7 mm Hg to about 4 mm Hg, or (2) *loss of blood volume,* which can decrease the mean systemic filling pressure to as little as zero.

"Pressure Gradient for Venous Return." When the right atrial pressure rises to equal the mean systemic filling pressure, all other pressures in the systemic circulation also approach this same pressure. Therefore, the pressures in all parts of the systemic circulation come closer and closer to the right atrial pressure. And, when all these pressures, including the right atrial

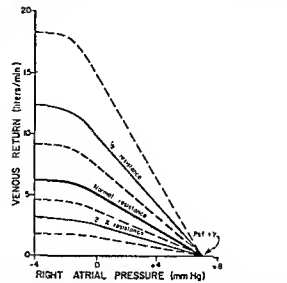


Figure 23–10. Venous return curves, depicting the effect of altering the resistance to venous return" (from Guyton, Jones, and Coleman Circulatory Physiology Cardiac Output and Its Regulation Philadelphia, W B Saunders Company, 1973)

pressure, finally reach the level of the mean systemic filling pressure, there is no longer any flow from any peripheral vessel back to the right atrium. However, when the right atrial pressure falls progressively lower than the mean systemic filling pressure, the flow to the heart increases proportionately, as one can see by studying any of the venous return curves in Figure 23–9. That is, *the greater the difference between the mean systemic filling pressure and the right atrial pressure the greater becomes the venous return* Therefore, the difference between these two pressures is called the *pressure gradient for venous return.*

Effect of Systemic Resistance on the Venous Return Curve. Figure 23–10 illustrates the effect of different systemic resistances on the venous return curve, showing that a decrease in resistance to one-half normal *rotates the curve upward,* while an increase in resistance

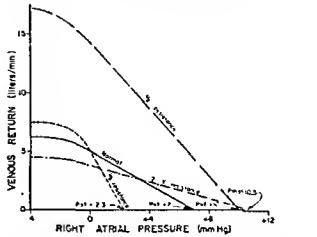


Figure 23–11. Combinations of the major patterns of venous return curves, illustrating the effects of simultaneous changes in mean systemic filling pressure *(Psf)* and in "resistance to venous return." (from Guyton, Jones, and Coleman Circulatory Physiology Cardiac Output and Its Regulation Philadelphia, W B Saunders Company, 1973)

rotates the curve downward. The venous return still becomes zero when the right atrial pressure rises to equal the mean systemic filling pressure, because when there is no pressure gradient to cause flow of blood, it makes no difference what the resistance is in the circulation—the flow is still zero.

Obviously, any factor that dilates the peripheral vessels decreases the resistance and rotates the function curve upward, while any factor that constricts the peripheral vessels rotates the curve downward. However, changes in venous resistance affect venous return about eight times as much as similar resistance changes in the arteries, because far more blood is stored in the veins than in the arteries

Combinations of Venous Return Curve Patterns. Figure 23–11 illustrates the effects on the venous return curve caused by simultaneous changes in mean systemic pressure and resistance to venous return, illustrating that both these factors can operate simultaneously

ANALYSIS OF CARDIAC OUTPUT AND RIGHT ATRIAL PRESSURE, USING CARDIAC OUTPUT AND VENOUS RETURN CURVES

In the complete circulation, the heart and the systemic circulation must operate together. This means (1) The venous return from the systemic circulation must equal the cardiac output from the heart (2) The right atrial pressure is the same for both the heart and for the systemic circulation

Therefore, one can predict the cardiac output and right atrial pressure in the following way (1) determine the momentary pumping ability of the heart and depict this in the form of a cardiac output curve, (2) determine the momentary state of the systemic circulation and depict this in the form of a venous return curve, and (3) "equate" these two curves against each other as shown in Figure 23–12.

The two black curves of Figure 23–12 depict both the *normal cardiac output curve* and the *normal venous return curve* Obviously, there is only one point on the graph, point A, at which the venous return equals the cardiac output and at which the right atrial pressure is the same in relation to both the heart and the systemic circulation Therefore, in the normal circulation the right atrial pressure, cardiac output, and venous return are all depicted by point A, called the *equilibrium point*

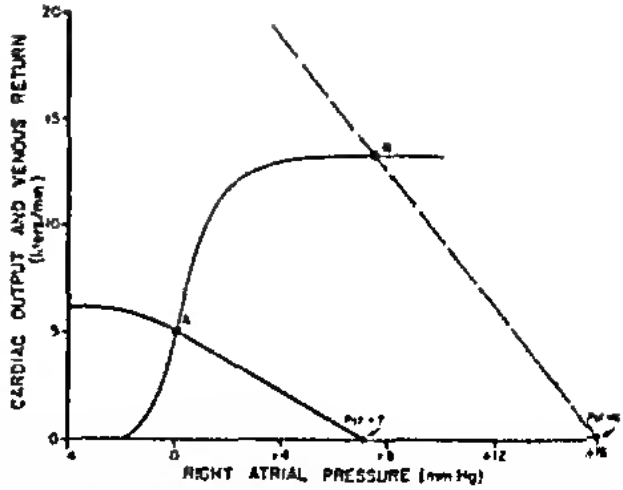


Figure 23–12. The two solid curves demonstrate an analysis of cardiac output and right atrial pressure when the cardiac output and venous return curves are known. Transfusion of blood equal to 20 per cent of the blood volume causes the venous return curve to become the dashed curve; as a result the cardiac output and right atrial pressure shift from point A to point B

Effect of Increased Blood Volume on Cardiac Output

A sudden increase in blood volume of about 20 per cent increases the cardiac output to about 2 5 to 3 times normal An analysis of this effect is illustrated by the dashed curve of Figure 23–12 Immediately upon infusing the large quantity of extra blood, the increased filling of the system causes the mean systemic filling pressure to rise to 16 mm Hg, which shifts the venous return curve upward and to the right At the same time, the increased blood volume distends the blood vessels, thus reducing their resistance and thereby rotating the curve upward As a result of these two effects, the venous return curve of Figure 23–12 changes from the solid curve to the dashed curve. This new curve equates with the cardiac output curve at point B, showing that the cardiac output increases 2 5 to 3 times and that the right atrial pressure rises to about +8 mm Hg

Compensatory Effects Initiated in Response to the Increased Blood Volume. The increased cardiac output caused by the increased blood volume lasts only a few minutes, because several different compensatory effects immediately begin to occur: (1) The increased cardiac output increases the capillary pressure so that fluid begins to transude out of the capillaries into the tissues, thereby returning the blood volume toward normal (2) The increased pressure in the veins causes the veins to distend gradually by the mechanism called *stress-relaxation*, especially causing the venous blood reservoirs such as the liver and spleen to distend (3) The excess blood flow through the peripheral tissues causes autoregulatory increase in the peripheral resistance These factors cause the mean systemic filling pressure to return back toward normal and also cause the resistance vessels of the systemic circulation to constrict. Therefore, gradually over a period of 20 to 40 minutes, the cardiac output returns almost to normal.

Effect of Sympathetic Stimulation on Cardiac Output

Sympathetic stimulation affects both the heart and the systemic circulation: (1) It makes the heart a stronger pump. (2) In the systemic circulation, it increases the mean systemic filling pressure because of contraction of the peripheral vessels—especially the veins, and it also increases the resistance to venous return The uppermost curves of Figure 23–13 depict analyses of the effects of moderate and maximal sympathetic stimulation on the cardiac output and right atrial pressure. The normal cardiac output and venous return curves are depicted by the very dark lines, moderate sympathetic stimulation is depicted by the long-dashed curves, and maximal sympathetic stimulation by the dot-dash colored curves

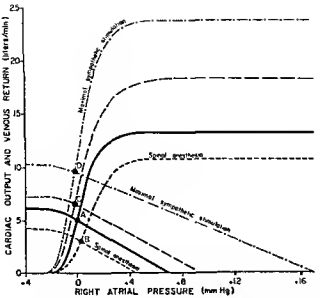


Figure 23–13. Analysis of the effect on cardiac output of (1) moderate sympathetic stimulation, (2) maximal sympathetic stimulation (colored curves), and (3) sympathetic inhibition caused by total spinal anesthesia. (From Guyton, Jones, and Coleman Circulatory Physiology: Cardiac Output and Its Regulation Philadelphia, W. B. Saunders Company, 1973.)

Note that maximal sympathetic stimulation (the colored curves) increases the mean systemic filling pressure (depicted by the point at which the venous return curve reaches the zero venous return level) to 17 mm Hg, and it also increases the pumping effectiveness of the heart by about 70 per cent. As a result, the cardiac output rises from the normal value at equilibrium point A to approximately double normal at equilibrium point D. Thus, different degrees of sympathetic stimulation can increase the cardiac output progressively to about two times normal—at least for short periods of time.

Effect of Sympathetic Inhibition

The sympathetic nervous system can be blocked completely by inducing *total spinal anesthesia* or by using some drug, such as *hexamethonium*, that blocks transmission of nerve impulses through the autonomic ganglia. The two lowermost curves in Figure 23–13, the short-dashed curves, show the effect of total sympathetic inhibition caused by total spinal anesthesia, illustrating that (1) the mean systemic filling pressure falls to about 4 mm Hg, and (2) the effectiveness of the heart as a pump decreases to about 80 per cent of normal. The cardiac output falls from point A to point B, which is a decrease to about 60 per cent of normal.

Increase in Cardiac Output in Exercise

Several different factors operate in exercise to increase the cardiac output. These include: (1) intense sympathetic stimulation, which strengthens the heart and also increases the mean systemic filling pressure, (2) contraction of the muscles around the blood vessels, which further increases the mean systemic filling pressure, and (3) intense dilatation of the resistance vessels in the muscles, thereby decreasing the resistance to venous return. Figure 23–14 depicts the effects of these factors on the circulation, showing that (1) compression of the blood vessels by the muscles plus sympathetic stimulation of the vessel walls increases the mean systemic filling pressure from 7 mm Hg to 20 mm Hg, (2) sympathetic stimulation increases the strength of the heart as a pump by about 70 per cent, and (3) the local vasodilatation in the muscles rotates the venous return curve upward to the right. As a result, the cardiac output and venous return curves in exercise (the dashed colored curves) equate at point B, which is a cardiac output of 22 liters per minute, or 4.5 times normal. The right atrial pressure hardly changes, indeed, when the degree of heart stimulation is great, the right atrial

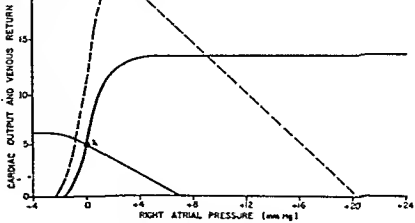


Figure 23–14. Graphical analysis of the changes in cardiac output and right atrial pressure with the onset of strenuous exercise.

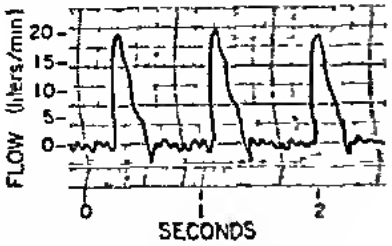


Figure 23–15. Pulsatile blood flow in the root of the aorta recorded by an electromagnetic flowmeter

pressure can sometimes fall to below normal despite a very high cardiac output

METHODS FOR MEASURING CARDIAC OUTPUT

In animal experiments, one can cannulate the aorta, the pulmonary artery, or the great veins entering the heart and measure the cardiac output using any type of flowmeter Also, an electromagnetic or ultrasonic flowmeter can be placed on the aorta or pulmonary artery to measure cardiac output However, except in rare instances, in the human being cardiac output is measured by indirect methods that do not require surgery. Two of the methods commonly used are the *oxygen Fick method* and the *indicator dilution method*.

Pulsatile Output of the Heart As Measured by an Electromagnetic or Ultrasonic Flowmeter

Figure 23–15 illustrates a recording in the dog of blood flow in the root of the aorta made using an electromagnetic flowmeter It demonstrates that the blood flow rises rapidly to a peak during systole and, then, at the end of systole, actually reverses for a small fraction of a second It is this reverse flow that causes the aortic valve to close And a minute amount of reverse flow continues throughout diastole to supply blood to the coronary vessels

Measurement of Cardiac Output by the Oxygen Fick Method

The Fick procedure is best explained by Figure 23–16 This figure shows that 200 ml of oxygen are being absorbed from the lungs into the pulmonary blood each minute. It also illustrates that the blood entering the right side of the heart has an oxygen concentration of 160 ml per liter of blood, whereas that leaving the left side has an oxygen concentration of 200 ml per liter of blood. From these data one can calculate that each liter of blood passing through the lungs picks up 40 ml of oxygen And, since the total quantity of oxygen absorbed into the blood from the lungs each minute is 200 ml, a total of 5 1-liter portions of blood must pass through the pulmonary circulation each minute to absorb this amount of oxygen. Therefore, the quantity of blood flowing through the lungs each minute is 5 liters, which is also a measure of the cardiac output. Thus, the cardiac output can be calculated by the following formula:

Cardiac output (liters/min) =

$$\frac{O_2 \text{ absorbed per minute by the lungs (ml/min)}}{\text{Arteriovenous } O_2 \text{ difference (ml/liter of blood)}}$$

19 - 14 = 5

In applying the Fick procedure, accurate determination of venous oxygen concentration can be achieved only by sampling blood directly from the right ventricle or preferably even the pulmonary artery, for blood in any one vein usually has a concentration of oxygen different from that in other veins, and even blood in the right atrium usually has not yet mixed satisfactorily. To obtain such a sample of mixed venous blood a catheter is usually inserted up the brachial vein of the forearm, through the subclavian vein, down to the right atrium, and finally into the right ventricle or pulmonary artery.

Blood used for determining the oxygen saturation in the arterial blood can be obtained from any artery in the body, because all arterial blood is thoroughly mixed before it leaves the heart and therefore has the same oxygen concentration.

The rate of oxygen absorption by the lungs is usually measured by a "respirometer," which will be described in Chapter 71. In essence, this device is a floating chamber containing oxygen that sinks in the water as oxygen is removed from it, thus measuring the rate of oxygen usage by the rate at which it falls.

THE INDICATOR DILUTION METHOD

In measuring the cardiac output by the indicator dilution method, a small amount of indicator, such as a dye, is injected into a large vein or preferably into the right side of the heart itself This passes rapidly through the right heart, the lungs, the left heart, and finally into the arterial system Then one records the concentration of the dye as it passes through one of the peripheral arteries, giving a curve such as one of the two colored curves illustrated in Figure 23–17. In each

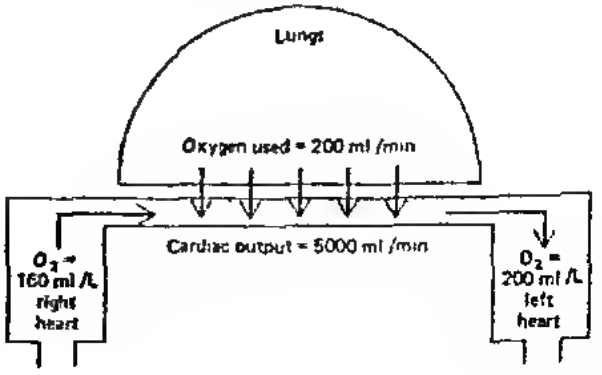


Figure 23–16. The Fick principle for determining cardiac output

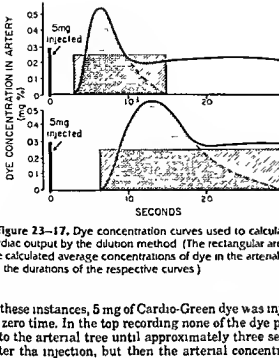


Figure 23–17. Dye concentration curves used to calculate the cardiac output by the dilution method (The rectangular areas are the calculated average concentrations of dye in the arterial blood for the durations of the respective curves)

of these instances, 5 mg of Cardio-Green dye was injected at zero time. In the top recording none of the dye passed into the arterial tree until approximately three seconds after tha injection, but then the arterial concentration of the dya rose rapidly to a maximum in approximately six to seven seconds After that, the concentration fell rapidly. However, before the concentration reached the zero point, some of the dye had already circulated all the way through some of the peripheral vessels and returned through the heart for a second time Consequently, the dya concentration in the artery began to rise again. For the purpose of calculation, however, it is necessary to extrapolate the early downslope of the curve to the zero point, as shown by the dashed portion of the curve. In this way, the *time-concentration curve* of the dye in an artery can be measured in its first portion and estimated reasonably accurately in its latter portion.'

Once the time-concentration curve has been determined, one can then calculate the mean concentration of dye in the arterial blood for the duration of the curve. In Figure 23–17, this was done by measuring the area under the entire curve, and then averaging the concentration of dye for the duration of the curve, one can see from the shaded rectangle straddling the upper curve of the figure that the average concentration of dye was approximately 0.25 mg/deciliter of blood and that the duration of this average value was 12 seconds. A total of 5 mg of dye was injected at the beginning of the experiment. In order for blood carrying only 0 25 mg of dye in each deciliter to carry the entire 5 mg of dye through the heart and lungs in 12 seconds, it would be necessary for a total of 20 1-deciliter portions of blood to pass through the heart during this time, which would be the same as a cardiac output of 2 liters per 12 seconds, or 10 liters per minute

In the bottom curve of Figure 23–17, the blood flow through the heart was considerably slower, and the dye did not appear in the arterial system until approximately six seconds after it had been injected. It reached a maximum height in 12 to 13 seconds and was extrapolated to 0 at approximately 30 seconds. Averaging the dye concentrations over the 24-second duration of the curve, one finds again an average concentration of 0 25 mg of dye in each deciliter of blood, but this time for a 24-second time interval instead of 12 seconds. To transport the total 5 mg of dye, 20 1-deciliter portions of blood would have had to pass through the heart during the 24-second time interval. Therefore, the cardiac output was 2 liters per 24 seconds, or 5 liters per minute.

To summarize, the cardiac output can be determined from the following formula

$$\text{Cardiac output (ml/min)} = \frac{\text{Milligrams of dye injected} \times 60}{\left(\begin{array}{c}\text{Average concentration of dye}\\ \text{in each milliliter of blood}\\ \text{for the duration of the curve}\end{array}\right) \times \left(\begin{array}{c}\text{Duration of}\\ \text{the curve}\\ \text{in seconds}\end{array}\right)}$$

Substances That Can Be Injected for Determining Cardiac Output by the Indicator Dilution Method. Almost any substance that can be analyzed satisfactorily in the arterial blood can be used as the "indicator" when making use of the indicator dilution method for determining cardiac output However, for optimum accuracy it is necessary that the injected substance not be lost into the tissues of the lungs during its passage to the sampling site. The most widely used substance is *Cardio-Green*, a dye that combines with the plasma proteins and, therefore, is not lost from the blood.

REFERENCES

Banet M and Guyton, A C Effect of body metabolism on cardiac output Role of the central nervous system *Am J Physiol*, *220* 662 1971

Bishop, V S and Stone, H L. Quantitative description of ventricular output curves in conscious dogs *Circ Res* *20* 581, 1967

Bishop, V S, *et al* Cardiac function curves in conscious dogs *Am J Physiol* *207* 677, 1964

Brecher, G A Venous Return New York Grune & Stratton, 1956

Bruce, T A, and Douglas, J E Dynamic cardiac performance *In* Frohlich, E D (ed) Pathophysiology 2nd Ed. Philadelphia, J B Lippincott Co, 1976, p 5

Coleman, T G, *et al* Control of cardiac output by regional blood flow distribution *Ann Biomed Eng*, 2 149, 1974

Dobbs, W A, *et al* Relative importance of nervous control of cardiac output and arterial pressure *Am J Cardiol* *27* 507, 1971

Dodge H T, and Kennedy, J W Cardiac output, cardiac performance, hypertrophy, dilatation, valvular disease, ischemic heart disease, and pericardial disease *In* Sodeman W A, Jr, and Sodeman, T M (eds) Pathologic Physiology Mechanisms of Disease 6th Ed Philadelphia, W B Saunders Co, 1979, p 271

Donald, D E, and Shepherd, J T Response to exercise in dogs with cardiac denervation *Am J Physiol* *205* 393, 1963

Dow, P Estimations of cardiac output and central blood volume by dye dilution *Physiol Rev*, *36* 77, 1956

Fermose, J D, *et al* Mechanism of decrease in cardiac output caused by opening the chest *Am J Physiol* *207* 1112 1964

Green, J F Determinants of systemic blood flow *In* Guyton, A C, and Young, D B (eds) International Review of Physiology Cardiovascular Physiology, III Vol 18 Baltimore University Park Press, 1979, p 33

Grodins, F S Integrative cardiovascular physiology A mathematical synthesis of cardiac and blood vessel hemodynamics *Q Rev Biol*, *34* 93, 1959

Guyton A G Determination of cardiac output by equating venous return curves with cardiac response curves *Physiol Rev*, *35* 123, 1955

Guyton, A G Venous return *In* Hamilton, W F (ed) Handbook of Physiology Sec. 2, Vol 2 Baltimore, Williams & Wilkins, 1963, p 1099

Guyton, A. G Regulation of cardiac output *N Engl J Med*, *277* 805, 1967

Guyton, A C An overall analysis of cardiovascular regulation *Anesth Analg*, *56* 761, 1977

Guyton, A C. Essential cardiovascular regulation—the control linkages between bodily needs and circulatory function *In* Dickinson, C. J.,

and Marks, J (eds) Developments in Cardiovascular Medicine Lancaster, England, MTP Press, 1978, p 265
Guyton, A. C, *et al*. Venous return at various right atrial pressures and the normal venous return curve *Am J Physiol*, *189* 609, 1957
Guyton, A C, *et al* Relative importance of venous and arterial resistance in controlling venous return and cardiac output *Am. J Physiol*, *196* 1008, 1959
Guyton, A C, *et al* Instantaneous increase in mean circulatory pressure and cardiac output at onset of muscular activity *Circ Res*, *11* 431, 1962
Guyton, A. C, *et al* Autoregulation of the total systemic circulation and its relation to control of cardiac output and arterial pressure *Circ Res 28 (Suppl 1)* 93 1971
Guyton, A C, *et al* Circulation Overall regulation *Annu Rev Physiol*, *34* 13 1972
Guyton, A C, *et al* Systems analysis of arterial pressure regulation and hypertension *Ann Biomed Eng 1* 254, 1972
Guyton A C, *et al* Cardiac Output and Its Regulation Philadelphia, W B Saunders Co, 1973
Ishikawa, N, *et al* Direct recording of cardiac output– and venous return–curves *Jap Heart J*, *19* 775 1978
Jones, C E *et al* Cardiac output and physiological mechanisms in circulatory shock *In* MTP International Review of Science Physiology Vol I Baltimore, University Park Press 1974, p 233
Kenner T (ed) Cardiovascular System Dynamics Models and Measurements New York, Plenum Publishing Corp, 1982
Keul, J The relationship between circulation and metabolism during exercise *Med. Sci Sports*, *5* 209 1973
Knoop A A Physiological aspects of circulatory dynamics especially related to ageing as studied by displacement ballistocardiography and other cardiovascular methods *Bibl Cardiol*, *30* 87 1973
Levy, M N and Zieske, H A closed circulatory system model *Physiologist 10* 419 1967
Longhurst, J C, and Mitchell, J H Reflex control of the circulation by afferents from skeletal muscle *In* Guyton, A C, and Young, D B (eds) International Review of Physiology: Cardiovascular Physiology III Vol 18 Baltimore, University Park Press, 1979, p 125
Mark, A L, and Mancia, G Cardiopulmonary baroreflexes in humans *In* Shepherd, J T, and Abboud, F M (eds) Handbook of Physiology Sec 2, Vol III Bethesda, American Physiological Society, 1983 p 795
Mitchell, J. H, and Wildenthal, K. Static (isometric) exercise and the heart. Physiological and clinical considerations *Annu Rev Med*, *25* 369, 1974
Nichols, W W Continuous cardiac output derived from the aortic pressure waveform A review of current methods *Biomed Eng*, *8* 376, 1973
Prather, J W, *et al* Effect of blood volume, mean circulatory pressure, and stress relaxation on cardiac output. *Am J Physiol*, *216* 467, 1969
Rothe, C F Reflex control of veins and vascular capacitance *Physiol Rev*, *63* 1281, 1983
Rothe, C F Venous system Physiology of the capacitance vessels *In* Shepherd, J T, and Abboud, F M (eds) Handbook of Physiology Sec 2, Vol III Bethesda, American Physiological Society, 1983, p 397
Sarnoff, S, and Mitchell J H The regulation of the performance of the heart *Am J Med*, *30* 747, 1961
Stone, H L, *et al* Ventricular function in cardiac-denervated and cardiac-sympathectomized conscious dogs *Circ Res*, *20* 587, 1967
Sugimoto, T, *et al* Effect of tachycardia on cardiac output during normal and increased venous return *Am J Physiol*, *211* 288, 1966
Varat, M A, *et al* Cardiovascular effects of anemia *Am Heart J*, *83* 415, 1972
Weisel, R. D, *et al* Current concepts measurement of cardiac output by thermodilution *N Engl J Med*, *292* 682 1975

24

The Pulmonary Circulation

The quantity of blood flowing through the lungs is essentially equal to that flowing through the systemic circulation. However, certain problems related to distribution of blood flow and other hemodynamics are special to the pulmonary circulation and are especially important to the gas exchange function of the lungs. Therefore, the present discussion is concerned specifically with the special features of blood flow in the pulmonary circuit and the function of the right side of the heart in maintaining this flow.

PHYSIOLOGIC ANATOMY OF THE PULMONARY CIRCULATORY SYSTEM

The Right Side of the Heart. As illustrated in Figure 24–1, the right ventricle is wrapped halfway around the left ventricle. The cause of this is the difference in pressures developed by the two ventricles during systole

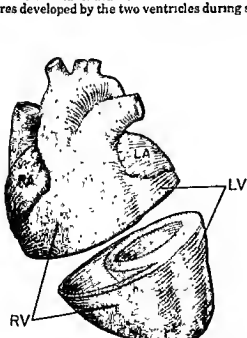


Figure 24–1. The anatomical relationship of the right ventricle to the left ventricle, showing the globular shape of the left ventricle and the half-moon shape of the right ventricle as it drapes around the left ventricle.

That is, the left ventricle contracts with extreme force in comparison with the right ventricle, which causes the left ventricle to assume a globular shape, and the septum protrudes into the right heart.

The muscle of the right ventricle is slightly more than one third as thick as that of the left ventricle; this also results from the difference in pressures between the two sides of the heart. Indeed, the wall of the right ventricle is only about three times as thick as the atrial walls, whereas the left ventricular muscle is about eight times as thick.

The Pulmonary Vessels. The pulmonary artery extends only 5 centimeters beyond the apex of the right ventricle and then divides into the right and left main branches, which supply blood to the two respective lungs. The pulmonary artery is also thin, with a wall thickness approximately twice that of the venae cavae and one third that of the aorta. The pulmonary arterial branches are all very short. However, all the pulmonary arteries, even the smaller arteries and arterioles, have much larger diameters than their counterpart systemic arteries. This, combined with the fact that the vessels are very thin and distensible, gives the pulmonary arterial tree a very large compliance, averaging about 2 ml per mm Hg, which is about half that of the entire systemic arterial tree. This large compliance allows the pulmonary arteries to accommodate most of the stroke volume output of the right ventricle.

The pulmonary veins, like the pulmonary arteries, are also short, but their distensibility characteristics are similar to those of the veins in the systemic circulation.

The Bronchial Vessels. Blood also flows to the lungs through several bronchial arteries, amounting to about 1 to 2 per cent of the total cardiac output. This bronchial arterial blood is *oxygenated* blood, in contrast to the partially deoxygenated blood in the pulmonary arteries. It supplies the supporting tissues of the lungs, including the connective tissue, the septa, and the large and small bronchi. After this bronchial arterial blood has passed through the supporting tissues, it empties into the pulmonary veins and *enters the left atrium* rather than passing back to the right atrium. Therefore, the left ventricular output is slightly greater than the right ventricular output.

The Lymphatics. Lymphatics extend from all the supportive tissues of the lung, beginning in the perivascular and peribronchial spaces and coursing to the hilum of the lung and thence mainly into the right lymphatic duct. Particulate matter entering the alveoli is usually

removed rapidly via these channels, and protein is also removed from the lung tissues, thereby preventing edema

PRESSURES IN THE PULMONARY SYSTEM

The Pressure Pulse Curve in the Right Ventricle. The pressure pulse curves of the right ventricle and pulmonary artery are illustrated in the lower portion of Figure 24–2. These are contrasted with the much higher aortic pressure curve shown above The systolic pressure in the right ventricle of the normal human being averages approximately 25 mm Hg, and the diastolic pressure averages about 0 to 1 mm Hg, values that are only one fifth those for the left ventricle

Pressures in the Pulmonary Artery. During systole, the pressure in the pulmonary artery is essentially equal to the pressure in the right ventricle, as also shown in Figure 24–2. However, after the pulmonary valve closes at the end of systole, the ventricular pressure falls precipitously, whereas the pulmonary arterial pressure falls slowly as blood flows through the capillaries of the lungs.

As shown in Figure 24–3, the systolic pulmonary arterial pressure averages approximately 25 mm Hg in the normal human being; the diastolic pulmonary arterial pressure, approximately 8 mm Hg; and the mean pulmonary arterial pressure, 15 mm Hg

Pulmonary Arterial Pulse Pressure. The pulse pressure in tha pulmonary arteries averages 17 mm Hg, which is two thirds as much as the systolic pressure. In previous discussions of pulse pressure in the systemic circulation, it has been pointed out that the less the compliance of an elastic reservoir that receives pulsatile injections of blood, the greater will be the pulse pressure. If it were not for the large compliance of the thin pulmonary arteries, the pulmonary arterial pulse

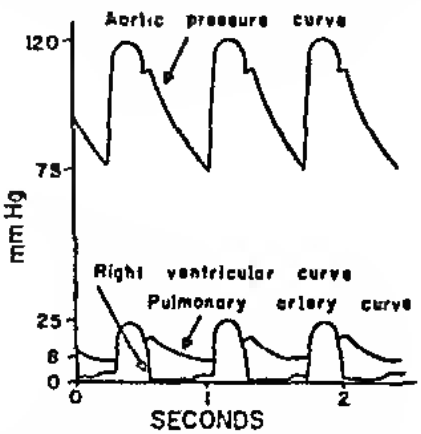


Figure 24–2. Pressure pulse contours in the right ventricle, pulmonary artery, and aorta.

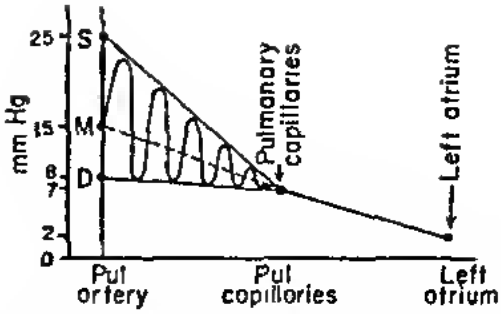


Figure 24–3. Pressures in the different vessels of the lungs.

pressure would be even greater than it is. However, another factor that keeps the pulse pressure from being higher is that about one half the blood ejected by the right ventricle runs off from the pulmonary arteries through the capillaries into the pulmonary veins and left atrium at the same time that it is being ejected during systole

Pulmonary Capillary Pressure. The mean pulmonary capillary pressure, as diagrammed in Figure 24–3, has been estimated by indirect means to be approximately 7 mm Hg. This will be discussed in detail later in the chapter in relation to fluid exchange functions of the capillary.

Left Atrial and Pulmonary Venous Pressure. The mean pressure in the left atrium and ia the major pulmonary veina averages approximately 2 mm Hg in the human being, varying from as low as 1 mm Hg to as high as 5 mm Hg.

It usually is not reasonable to measure tha left atrial pressure directly in the normal human being because it is difficult to pass a catheter through the heart chambers into the left atrium. However, the left atrial pressure often can be estimated closely by measuring the so-called *pulmonary wedge pressure* This is achieved by inserting a catheter through the right heart and pulmonary artery into one of the small branch pulmonary arteries and then pushing the catheter until it wedges tightly in the artery. The pressure then measured through the catheter, called the "wedge pressure," is about 5 mm Hg. Since all blood flow has been stopped in the small artery and the blood vessels extending from the artery make almost direct connection through the pulmonary capillaries with the blood in the pulmonary veins, this wedge pressure is usually only 2 to 3 mm Hg greater than the left atrial pressure. And when the left atrial pressure rises to high values, so also rises the pulmonary wedge pressure. Wedge pressure measurements are used frequently for studying changes in left atrial pressure in congestive heart failure.

THE BLOOD VOLUME OF THE LUNGS

The blood volume of the lungs is approximately 450 ml, about 9 per cent of the total blood volume of the circulatory system About 70 ml of this is in

the capillaries, and the remainder is divided about equally between the arteries and veins.

The Lungs as a Blood Reservoir. Under different physiological and pathological conditions, the quantity of blood in the lungs can vary from as little as 50 per cent of normal up to as high as 200 per cent of normal For instance, when a person blows air out so hard that high pressure is built up in the lungs—such as when blowing a trumpet—as much as 250 ml of blood can be expelled from the pulmonary circulatory system into the systemic circulation. Also, loss of blood from the systemic circulation by hemorrhage can be partly compensated for by automatic shift of blood from the lungs into the systemic vessels

Shift of Blood Between the Pulmonary and Systemic Circulatory Systems as a Result of Cardiac Pathology. Failure of the left side of the heart or increased resistance to blood flow through the mitral valve as a result of mitral stenosis or mitral regurgitation causes blood to dam up in the pulmonary circulation, sometimes increasing the pulmonary blood volume as much as 100 per cent, and also causing corresponding increases in the pulmonary vascular pressures

On the other hand, exactly the opposite effects take place when the right side of the heart fails.

Because the volume of the systemic circulation is about nine times that of the pulmonary system, a shift of blood from one system to the other affects the pulmonary system greatly but usually has only mild systemic effects.

BLOOD FLOW THROUGH THE LUNGS AND ITS DISTRIBUTION

The blood flow through the lungs is essentially equal to the cardiac output. Therefore, the factors that control cardiac output—mainly peripheral factors, as discussed in Chapter 23—also control pulmonary blood flow. Under most conditions, the pulmonary vessels act as passive, distensible tubes that enlarge with increasing pressure and narrow with decreasing pressure But, for adequate aeration of the blood it is important for the blood to be distributed to those segments of the lungs where the alveoli are best oxygenated. This is achieved by the following mechanism:

Effect of Diminished Alveolar Oxygen on Local Alveolar Blood Flow—Automatic Control of Pulmonary Blood Flow Distribution. When the concentration of oxygen in some alveoli becomes very low, the adjacent blood vessels slowly constrict during the ensuing three to ten minutes, the vascular resistance increasing to as much as triple normal. It should be noted specifically that this is *opposite to the effect* normally observed in systemic vessels, which dilate rather than constrict in response to low oxygen. However, it is believed that the low oxygen concentration causes some vasoconstrictor substance to be released from the lung tissue, this substance in turn promoting small arterial and arteriolar constriction. Unfortunately, research workers have not yet been able to isolate the vasoconstrictor substance.

This effect of low oxygen on pulmonary vascular resistance has an important function: to distribute blood flow where it is most effective. That is, when some of the alveoli are poorly ventilated so that the oxygen concentration in them becomes low, the local vessels constrict. This in turn causes most of the blood to flow through the other areas of the lungs that are better aerated, thus providing an automatic control system for distributing blood flow through different pulmonary areas in proportion to their degrees of ventilation.

Paucity of Autonomic Nervous Control of Blood Flow in the Lungs. Though nerves innervate the lung tissues, it is doubtful that these have a major function in normal control of pulmonary blood flow. Normally, stimulation of the vagal fibers to the lungs causes a slight decrease in pulmonary vascular resistance, and stimulation of the sympathetics causes a slight increase in resistance, both effects perhaps too little to be more than marginally important.

However, research workers have described reflexes in the pulmonary vascular system that might be of clinical importance under certain conditions. For instance, it has been claimed that small emboli occluding the small pulmonary arteries cause a reflex that elicits sympathetic vasoconstriction throughout the lungs, this vasoconstriction then leading to an intense increase in pulmonary arterial pressure However, the significance of this reflex is still not certain

EFFECT OF HYDROSTATIC PRESSURE GRADIENTS IN THE LUNGS ON REGIONAL PULMONARY BLOOD FLOW

In Chapter 19 it was pointed out that the pressure in the foot of a standing person can be as much as 90 mm Hg greater than the pressure at the level of the heart This is caused by *hydrostatic pressure*—that is, by the weight of the blood itself The same effect, but to a lesser degree, occurs in the lungs In the normal, upright adult person, the lowest point in the lungs is about 30 cm below the highest point. This represents a 23 mm Hg pressure difference, about 15 mm Hg of which are above the heart and 8 below That is, the pulmonary arterial pressure in the uppermost portion of the lung of a standing person is about 15 mm Hg less than the mean pulmonary arterial pressure, and the pressure in the lowest portion of the lungs is about 8 mm Hg greater. Such pressure differences have profound effects on blood flow through the different areas of the lungs. This is illustrated by the lower curve in Figure 24–4, which plots blood flow per unit of lung tissue against hydrostatic level in the lung Note that in the standing position there is little flow in the top of the lung but three to four times this much flow in the lower lung. To help explain the reasons for these differences, the lung is often divided into three different zones, as illustrated

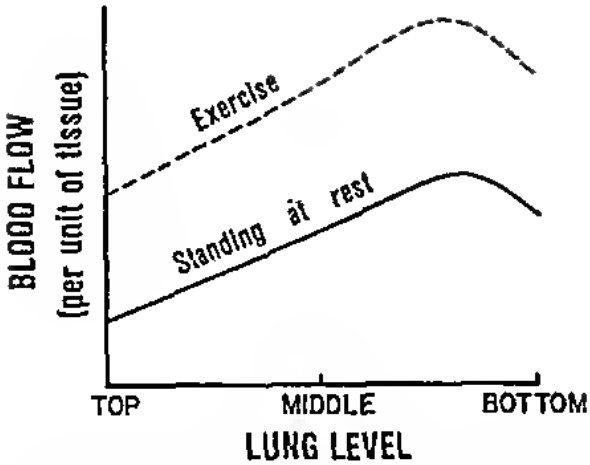


Figure 24–4. Blood flow at different levels in the lung of an upright person, both at rest and during exercise. Note that when the person is at rest, the blood flow is almost zero at the top of the lungs and most of the flow is through the lower lung.

in Figure 24–5, in which the patterns of blood flow are quite different. Let us explain these differences.

Zone 1 (Area of Zero Flow). When the pulmonary arterial pressure is very low and the person is in a standing position, the pressure in the capillaries at the apices of the lungs may be less than the pulmonary alveolar pressure (which normally averages 0 mm Hg). In such a condition, the capillaries remain collapsed all the time, as illustrated in the top panel of Figure 24–5, because the pulmonary capillary pressure is not great enough to oppose the tendency of the alveolar pressure to collapse the capillaries. Therefore, there will be zero blood flow. Thus, *zone 1* is an area of no flow.

Note, however, that a zone 1 area *does not occur in the normal lung* but instead occurs only in abnormal conditions.

Zone 2 (Area of Zero Flow During Part of the Heart Cycle). Recall that the normal pulmonary arterial pressure ranges between 8 mm Hg during diastole and 25 mm Hg during systole. Since some portions of the normal upright lung are at a hydrostatic pressure level as much as 15 mm Hg above the mean level of the heart, during diastole the pulmonary arterial pressure cannot force blood into the upper portions of the normal lung. Yet, during systole blood does flow. Therefore, as illustrated in the middle panel of Figure 24–5, the alveolar capillaries collapse during the diastolic portion of the heart cycle but open during systole. Thus, in *zone 2*, blood flows intermittently, that is, with zero flow during part of the heart cycle.

In the normal lung under normal standing conditions, the lower part of zone 2 begins 7 to 10 cm above the level of the heart and extends from there all the way to the top of the lungs.

Zone 3 (Area of Continuous Flow). In the lower part of the lung, the pulmonary vascular pressures are always higher than the alveolar pressure, even during diastole. Therefore, as illustrated in the lower part of Figure 24–5, the alveolar capillaries remain distended all the time, and blood flows continuously from the arteries to the veins.

The zone 3 portion of the lung, the portion in which the pulmonary capillaries remain open all the time, normally begins 7 to 10 cm above the level of the heart and extends all the way to the lowermost portions of the lung.

Factors That Can Change the Locations of Zones 1, 2, and 3 in the Lungs. Though normally there is no zone 1 blood flow in the lungs, this very often does occur when the pulmonary arterial pressure falls. This is very common following hemorrhage. However, the most common cause of a zone 1 area in the lungs is breathing against a high air pressure, as occurs when a person is blowing on a musical instrument. Under these conditions, the alveolar pressure rises very high, so high that the pulmonary arterial systolic pressure in the upper part of the lung cannot overcome the compressive effect on the capillaries caused by the high alveolar pressure. Consequently, no blood then flows through this portion of the lungs.

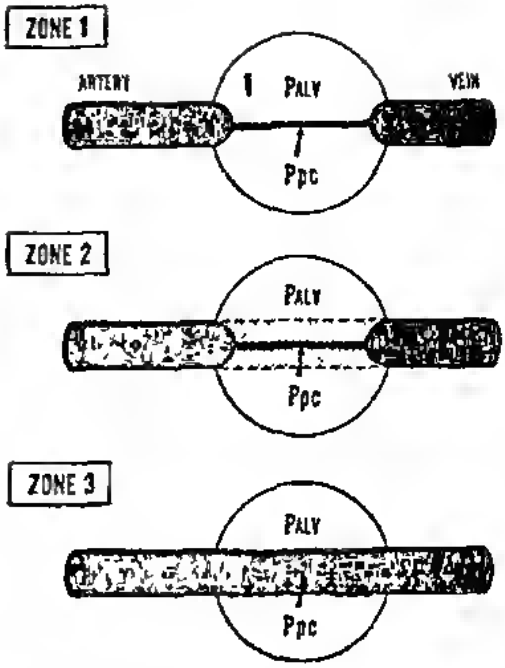


Figure 24–5. Mechanics of blood flow in the three different blood flow zones of the lung. *zone 1*, no flow, because alveolar pressure is greater than arterial pressure; zone 2, intermittent flow, because systolic arterial pressure rises higher than alveolar pressure but diastolic pressure falls below alveolar pressure. *zone 3*, continuous flow, because arterial pressure remains greater than alveolar pressure at all times.

Another factor that affects the zones is position of the body. In the lying position, all portions of the lung become zone 3, and blood flow in the different parts of the lung then becomes rather evenly distributed. In the upside-down position, blood flow to the apices can actually be greater than blood flow to the base of the lung.

Effect of Exercise on Blood Flow Through the Lungs. Referring again to Figure 24–4, one sees that the blood flow in all parts of the lung increases during exercise. However, the increase in flow in the top of the lung may be as much as 700 to 800 per cent, whereas the increase in the lower part of the lung may be no more than 200 to 300 per cent. The reason for these effects is the considerably higher pulmonary vascular pressures that occur during exercise, which effectively convert the entire lung into a zone 3 pattern of flow.

EFFECT OF INCREASED CARDIAC OUTPUT ON THE PULMONARY CIRCULATION DURING HEAVY EXERCISE

During heavy exercise the lungs are frequently called upon to absorb up to 20 times as much oxygen into the

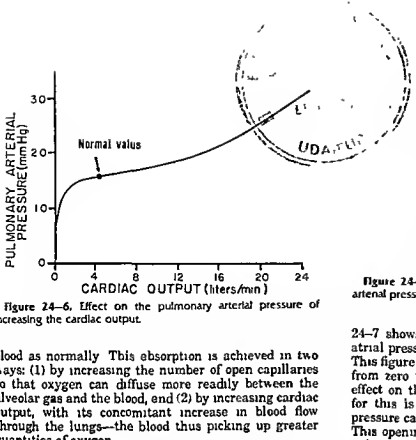


Figure 24–6. Effect on the pulmonary arterial pressure of increasing the cardiac output.

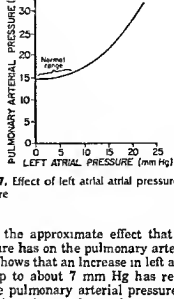


Figure 24–7. Effect of left atrial atrial pressure on pulmonary artenal pressure

blood as normally This absorption is achieved in two ways: (1) by increasing the number of open capillaries so that oxygen can diffuse more readily between the alveolar gas and the blood, and (2) by increasing cardiac output, with its concomitant increase in blood flow through the lungs—the blood thus picking up greater quantities of oxygen

Fortunately, the cardiac output can increase to four to six times normal before pulmonary arterial pressure becomes excessively elevated, this effect is illustrated in Figure 24–6. As the blood flow into the lungs becomes increased, more and more capillaries open up; also, the pulmonary arterioles and capillaries expend Therefore, the excess flow passes through the pulmonary vascular system without excessive increase in pulmonary arterial pressure. Indeed, as shown in the figure, the pressure rarely rises more than twofold despite as much as a five- to sixfold increase in flow.

This ability of the lungs to accommodate greatly increased blood flow during exercise with relatively little increase in pulmonary vascular pressure is important for at least two reasons (1) it obviously conserves the energy of the right heart, and (2) it prevents a significant rise in pulmonary capillary pressure and therefore also prevents development of pulmonary edema during the increased cardiac output

FUNCTION OF THE PULMONARY CIRCULATION WHEN THE LEFT ATRIAL PRESSURE RISES AS A RESULT OF LEFT HEART FAILURE

When the left heart fails or when mitral valvular disease blocks the outflow of blood from the left atrium, the left atrial pressure is increased and blood becomes dammed up in the pulmonary system The initial effect is to increase the volume of blood in the pulmonary veins. Then, with further increase in left atrial pressure, the capillaries and the pulmonary arteries also become excessively filled with blood, and thus damming of blood in the pulmonary arteries eventually raises the pulmonary arterial pressure. Also, an important effect of the rising capillary pressure is transudation of fluid into the pulmonary tissues and alveoli, an event called *pulmonary edema*, which will be discussed subsequently.

Quantitative Relationship Between Left Atrial Pressure and Pulmonary Arterial Pressure. Figure 24–7 shows the approximate effect that a rising left atrial pressure has on the pulmonary arterial pressure This figure shows that an increase in left atrial pressure from zero up to about 7 mm Hg has relatively little effect on the pulmonary arterial pressure The reason for this is that this moderate increase in left atrial pressure causes more and more opening up of capillaries This opening of blood vessels progressively reduces the pulmonary vascular resistance, which mainly compensates for the back pressure effect of the rising left atrial pressure. However, once all these vessels have been opened, any further rise in left atrial pressure above the 7 mm Hg level causes a marked rise in pulmonary arterial pressure, as shown in the figure

So long as the left ventricle functions normally, the left atrial pressure rarely rises above 6 mm Hg—even when it is pumping maximum cardiac output Therefore, the left atrial pressure normally does not rise high enough to have any significant effect on pulmonary circulatory function Yet, when the left heart fails, the left atrial pressure does then rise very high, sometimes to as high as 30 to 40 mm Hg, which often leads to a serious increase in pulmonary arterial pressure with concomitant excess load on the right heart.

PULMONARY CAPILLARY DYNAMICS

Exchange of gases between the alveolar air and the pulmonary capillary blood will be discussed in Chapter 40. However, it is important for us to note here that the alveolar walls are lined with so many capillaries that in most places they almost touch each other. Therefore, it has often been said that the capillary blood flows in the alveolar walls as a "sheet" rather than in individual vessels.

Pulmonary Capillary Pressure. Unfortunately, no direct measurements of pulmonary capillary pressure have been made. However, "isogravimetric" measurement of pulmonary capillary pressure, using a technique described in Chapter 30, has given a value of 7 mm Hg. This is probably very nearly correct, because the mean left atrial pressure is about 2 mm Hg and the mean pulmonary arterial pressure is only 15 mm Hg, so that the mean pulmonary capillary pressure must lie somewhere between these two values.

Length of Time Blood Stays in the Capillaries. From histologic study of the total cross-sectional area of all the pulmonary capillaries, it can be calculated that when the cardiac output is normal, blood passes through the pulmonary capillaries in about 0.8 second. Increasing the cardiac output shortens this time sometimes to as little as 0.3 second, but the shortening would be much greater were it not for the fact that additional capillaries, which normally are collapsed, open up to accommodate the increased blood flow. Thus, in less than 1 second, blood passing through the capillaries becomes oxygenated and loses its excess carbon dioxide.

CAPILLARY EXCHANGE OF FLUID IN THE LUNGS

Negative Interstitial Fluid Pressure in the Lungs and Its Significance. Fluid exchange through the capillary membranes of the systemic circulation will be discussed in detail in Chapter 30, where it is pointed out that the most significant difference between pulmonary capillary dynamics and capillary dynamics elsewhere in the body is the very low capillary pressure in the lungs, about 7 mm Hg, in comparison with a considerably higher "functional" capillary pressure elsewhere in the body, probably about 17 mm Hg. Because of the very low pulmonary capillary pressure, the hydrostatic force tending to push fluid out the capillary pores into the interstitial spaces is also very slight. Yet, the colloid osmotic pressure of the plasma, about 28 mm Hg, is a large force tending to pull fluid into the capillaries. Therefore, there is continual osmotic tendency to dehydrate the interstitial spaces of the lungs. Referring to Chapter 30 again, it is calculated there that the normal pulmonary interstitial fluid pressure is probably about −8 mm Hg. That is, there is approximately 8 mm Hg tending to pull the alveolar epithelial membrane toward the capillary membrane, thus squeezing the pulmonary interstitial space down to almost nothing. Electron micrographic studies have demonstrated this fact, the interstitial space at times being so narrow that the basement membrane of the alveolar epithelium is fused with the basement membrane of the capillary endothelium. As a result, the distance between the air in the alveoli and the blood in the capillaries is minimal, averaging about 0.4 micron in distance; this obviously allows very rapid diffusion of oxygen and carbon dioxide.

The details of fluid exchange through the pulmonary capillary membrane will be discussed in relation to overall capillary function in Chapter 30.

Mechanism by Which the Alveoli Remain Dry. Another consequence of the negative pressure in the interstitial spaces is that it pulls fluid from the alveoli through the alveolar membrane and the interstitial spaces, thereby normally keeping the alveoli dry.

Pulmonary Edema. Pulmonary edema means excessive quantities of fluid either in the pulmonary interstitial spaces or in the alveoli. The normal pulmonary interstitial fluid volume is approximately 20 per cent of the lung mass, but this can increase perhaps to 30 per cent in pulmonary edema. In addition, many times this much fluid can enter the alveoli and cause *intra-alveolar edema.* Intra-alveolar fluid sometimes reaches 500 to 1000 per cent of the normal interstitial fluid volume.

Safety Factor Against Pulmonary Edema. The most common cause of pulmonary edema is greatly elevated pulmonary capillary pressure resulting from failure of the left heart and consequent damming of blood in the lungs. However, the pulmonary capillary pressure usually must rise to very high values before serious pulmonary edema will develop. The reason for this is the very high dehydrating force of the colloid osmotic pressure of the blood in the lungs. This effect is illustrated in Figure 24–8, which shows development of lung edema in dogs subjected to progressively increasing left atrial pressure. In this experiment no edema fluid developed in the lungs until left atrial pressure rose above 23 mm Hg, which was approximately 3 mm Hg greater than the colloid osmotic pressure of dog blood, about 20 mm Hg. This experiment demonstrates that the blood pressure in the pulmonary capillaries usually must rise a few mm Hg above the colloid osmotic pressure before serious pulmonary edema can ensue. In the human being the colloid osmotic pressure is about 28 mm Hg, so that pulmonary edema will rarely develop below 30 mm Hg pulmonary capillary pressure. Thus, if the capillary pressure in the lungs is normally 7 mm Hg and this pressure must usually rise above 30 mm Hg before edema will occur, the lungs have a *safety factor against edema* of approximately 23 mm Hg.

Safety Factor Against Edema in Chronic Elevation of Capillary Pressure. Patients with chronic elevation of pulmonary capillary pressure occasionally will not develop pulmonary edema even with pulmonary capillary pressures as high as 45 mm Hg. The reason for this is probably extremely rapid run-off of fluid from the pulmonary interstitial spaces through the lymphatics. In fact, experiments in dogs have shown that when the pulmonary capillary pressure remains elevated for

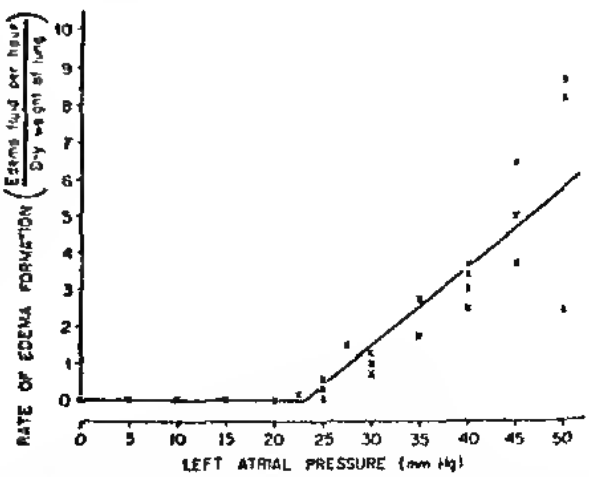


Figure 24–8. Rate of fluid loss into the lung tissues when the left atrial pressure (and also pulmonary capillary pressure) is increased (from Guyton and Lindsey: *Circ. Res.*, 7:649, 1959. By permission of The American Heart Association, Inc.)

more than two weeks, the pulmonary lymphatics can enlarge so much that lymph flow increases as much as twentyfold above the normal resting level. This extra lymph flow gives a safety factor perhaps 15 or more mm Hg above that which one normally has against edema.

Pulmonary Edema as a Result of Capillary Damage. Pulmonary edema can also result from local capillary damage in the lungs This effect is often caused by bacterial infection, such as occurs in pneumonia, or by irritant gases such as chlorine, sulfur dioxide, or war gases—mustard gas, for instance All these directly damage the alveolar epithelium and the endothelium of the capillaries, allowing rapid transudation of both fluid and protein into the alveoli and interstitial spaces

Alveolar Fluid in Pulmonary Edema. Though mild degrees of pulmonary edema can be limited to an increase only of the pulmonary interstitial fluid, serious edema almost always causes fluid transudation into the alveoli themselves It seems that the alveolar epithelium does not have enough strength to resist any significant degree of positive pressure in the interstitial spaces Therefore, minute or large ruptures occur in the epithelium, and fluid flows readily out of the interstitial spaces into the alveoli.

PATHOLOGICAL CONDITIONS THAT OBSTRUCT BLOOD FLOW THROUGH THE LUNGS

Removal of Lung Tissue. When one *entire lung* is removed, the flow of blood through the remaining lung usually is well within the limits of compensation as long as the patient remains inactive However, the patient thereafter has far less *pulmonary circulatory reserve* than does the normal person, for if the cardiac output increases to more than 150 to 200 per cent above normal, the pulmonary arterial pressure begins to rise rapidly This compares with the normal individual whose pulmonary circulatory reserve is 300 to 500 per cent above normal before the pressure rises seriously

Massive Pulmonary Embolism. One of the most severe postoperative calamities in surgical practice is massive pulmonary embolism Patients lying immobile in bed tend to develop extensive clots, especially in the veins of the legs, because of sluggish blood flow Also, women, after delivery of their babies, frequently develop massive clots in the hypogastric veins Such clots often break away from the initial sites of formation, particularly when the patient first walks after a long period of immobilization. The clots then flow to the right side of the heart and into the pulmonary artery Such a freemoving clot is called an *embolus*

Total blockage of only one of the major branches of the pulmonary artery usually is not immediately fatal because the opposite lung can accommodate all the blood flow. However, blood clots, as was discussed in Chapter 9, have a tendency to grow. Consequently, the embolus becomes larger and larger, and, as it extends into the other major branch of the pulmonary artery, the few remaining vessels that do not become plugged are taxed beyond their limit, and death ensues because of an inordinate rise in pulmonary arterial pressure and right-sided heart failure.

If the blood is rendered less coagulable by administration of anticoagulants so that the clot cannot grow, the life of the patient can often be saved, also, surgical removal of the clot is sometimes successful.

Diffuse Pulmonary Embolism. Sometimes many small blood clot emboli or small emboli of other substances besides blood clots enter the venous blood and eventually clog the vessels of the lungs. One such instance is *fat embolism.* This occurs when a large volume of fatty tissue, such as that in the breast, becomes traumatized or infected so that the fatty material liquefies and enters the venous blood. It passes to the lungs, and because of the high surface tension of fat, the pulmonary arterial pressure often cannot force the fat globules through the small capillary vessels Consequently, many of the small vessels of the lungs become blocked

Another instance is *air embolism* The air, like fat, has a large surface tension at the interface between the blood and air so that the globules of air cannot be deformed enough to be pushed through the capillaries.

The physical effects of diffuse pulmonary embolism on the pulmonary circulatory system are similar to those of massive pulmonary embolism—that is, increased pulmonary circulatory resistance with resulting increase in pulmonary arterial pressure and failure of the right side of the heart. However, with diffuse embolism, the pulmonary vessels appear to develop considerable vasospasm which adds additional resistance to flow besides that caused by the emboli themselves It is believed that this vasospasm is caused by a sympathetic reflex, and, clinically, the stellate ganglion is sometimes anesthetized to block the reflex.

Patients with this type of embolism exhibit a rapid respiratory rate because of local irritation by the emboli in the lungs and also because of resultant ischemia throughout the body.

Emphysema. Pulmonary emphysema means literally too much air in the lungs, and it is usually characterized by destruction of many of the alveolar walls This causes the adjacent alveoli to become confluent, thereby forming large *emphysematous cavities* rather than the many much smaller alveoli Obviously, loss of the alveolar septa greatly decreases the total alveolar surface area of the lungs and hinders gas exchange between the alveoli and the blood. This effect will be discussed in connection with the gas exchange functions of the lungs in Chapter 43

Emphysema also has an important effect on the pulmonary vasculature, for each time an alveolar wall is destroyed, some of the small blood vessels of the pulmonary system are also destroyed, thus progressively increasing the pulmonary resistance and elevating the pulmonary arterial pressure

In addition, a *physiologic increase in vascular resistance* also occurs in emphysema in the following way: Some of the emphysematous alveoli have poor exchange of air with the atmosphere. This poor ventilation results in alveolar hypoxia so that the adjacent blood vessels come continually under the influence of the hypoxic stimulus discussed earlier in the chapter, resulting in chronic vasoconstriction. This compounds the difficulties already caused by the pathological destruction of pulmonary vessels. Furthermore, the generalized hypoxia throughout the body causes increased cardiac output, and this, combined with the increased pulmonary resistance, sometimes results in serious pulmonary hypertension and right heart failure.

The physiologic increase in resistance can often be treated effectively by simple oxygen therapy. Therefore, emphysematous patients who also have right heart failure often experience rapid reversal of the failure

along with decrease of the dyspneic symptoms after only a short period of oxygen breathing.

Unfortunately, the prevalence of emphysema is increasing rapidly because of cigarette smoking. Therefore, the physiological principles of its treatment are becoming progressively more important.

Diffuse Sclerosis of the Lungs. A number of pathological conditions cause excessive fibrosis in the supportive tissues in the lungs, and the fibrous tissue in turn contracts around the vessels. Some of these conditions are silicosis, tuberculosis, syphilis, and, to a lesser extent, anthracosis.

In early stages of diffuse sclerosis, the pulmonary arterial pressure often is normal as long as the person is not exercising, but just as soon as even mild exercise is performed, the pulmonary arterial pressure rises inordinately because the vessels do not have the ability to expand as much as normal pulmonary vessels do. In late stages of diffuse sclerosis the pulmonary arterial pressure remains elevated constantly; as a result, the right ventricle hypertrophies, and it may fail. Diffusion of gases through the alveolar membranes in most instances is impaired either because of decreased surface area or because of diminished blood flow to some alveoli.

Atelectasis. Atelectasis is the clinical term for collapse of a lung or part of a lung. This occurs often when the bronchi become plugged, because the pulmonary blood rapidly absorbs the air in the entrapped alveoli, which causes the alveoli to collapse.

Atelectasis also occurs when the chest cavity is opened to atmospheric pressure, for, when air is allowed to enter the pleural space, the elastic nature of the lungs causes them to collapse immediately.

When the elastic tissues of the lungs contract during atelectasis, they constrict not only the alveoli but also the blood vessels. This automatically decreases blood flow in the atelectatic portions of the lungs to perhaps as little as one-fourth normal, shifting the remaining three fourths to the aerated portions. This is an important safety mechanism, for it prevents flow of major quantities of blood through collapsed, nonaerated pulmonary areas.

REFERENCES

Bergofsky, E. H. Humoral control of the pulmonary circulation. *Annu. Rev. Physiol.* 42:221, 1980.

Cournand, A. Some aspects of the pulmonary circulation in normal man and in chronic cardiopulmonary diseases. *Circulation*, 2:641, 1950.

Culver, B. H., and Butler, J. Mechanical influences on the pulmonary microcirculation. *Annu. Rev. Physiol.*, 42:187, 1980.

Cumming, G. The pulmonary circulation. *In* MTP International Review of Science: Physiology. Baltimore, University Park Press. Vol. 1, 1974, p. 93.

Dawson, C. A. Role of pulmonary vasomotion in physiology of the lung. *Physiol. Rev.*, 64:544, 1984.

Downing, S. E., and Lee, J. C. Nervous control of the pulmonary circulation. *Annu. Rev. Physiol.*, 42:199, 1980.

Effros, R. M. Pulmonary microcirculation and exchange. *In* Renkin, E. M., and Michel, C. C. (eds.) Handbook of Physiology. Sec. 2, Vol. IV. Bethesda, American Physiological Society, 1984, p. 865.

Fishman, A. P. Hypoxia in the pulmonary circulation. *Circ. Res.*, 38:221, 1976.

Fishman, A. P. Vasomotor regulation of the pulmonary circulation. *Annu. Rev. Physiol.*, 42:211, 1980.

Fishman, A. P., and Pietra, G. G. Hemodynamic pulmonary edema. *In* Fishman, A. P., and Renkin, E. M. (eds.) Pulmonary Edema. Baltimore, Waverly Press, 1979, p. 79.

Gaar, K. A., *et al.* Pulmonary capillary pressure and filtration coefficient in the isolated perfused lung. *Am. J. Physiol.*, 213:910, 1967.

Gil, J. Influence of surface forces on pulmonary circulation. *In* Fishman, A. P., and Renkin, E. M. (eds.) Pulmonary Edema. Baltimore, Waverly Press, 1979, p. 53.

Gil, J. Organization of microcirculation in the lung. *Annu. Rev. Physiol.*, 42:177, 1980.

Grover, R. F., *et al.* Pulmonary Circulation. *In* Shepherd, J. T., and Abboud, F. M. (eds.) Handbook of Physiology. Sec. 2, Vol. III. Bethesda, American Physiological Society, 1983, p. 103.

Guyton, A. C. Introduction to Part I. Pulmonary alveolar-capillary interface and interstitium. *In* Fishman, A. P., and Hecht, H. H. (eds.) The Pulmonary Circulation and Interstitial Space. Chicago, University of Chicago Press, 1969, p. 3.

Guyton, A. C., *et al.* Dynamics of subatmospheric pressure in the pulmonary interstitial fluid. *In* Lung Liquids. Ciba Symposium. New York: Elsevier North-Holland, 1976, p. 77.

Guyton, A. C., *et al.* Forces governing water movement in the lung. *In* Pulmonary Edema. Washington, D.C., American Physiological Society, 1979, p. 65.

Hasian, J. C., *et al.* Pressure-volume curves of systemic and pulmonary circuits. *Am. J. Physiol.* 213:1499, 1967.

Hughes, J. M. B. Pulmonary circulatory and fluid balance. *Int. Rev. Physiol.*, 14:135, 1977.

Malik, A. B. Pulmonary microembolism. *Physiol. Rev.*, 63:1114, 1983.

Mark, A. L., and Mancia, G. Cardiopulmonary baroreflexes in humans. *In* Shepherd, J. T., and Abboud, F. M. (eds.) Handbook of Physiology. Sec. 2, Vol. III. Bethesda: American Physiological Society, 1983, p. 795.

Meyer, B. J., *et al.* Interstitial fluid pressure. V. Negative pressure in the lungs. *Circ. Res.* 22:263, 1968.

Parker, J. C., *et al.* Pulmonary interstitial and capillary pressures estimated from intra-alveolar fluid pressures. *J. Appl. Physiol.*, 44(2):267, 1978.

Parker, J. C., *et al.* Pulmonary transcapillary exchange and pulmonary edema. *In* Guyton, A. C., and Young, D. B. (eds.) International Review of Physiology: Cardiovascular Physiology III, Vol. 18. Baltimore, University Park Press, 1979, p. 261.

Permutt, S., *et al.* Effect of lung inflation on static pressure-volume characteristics of pulmonary vessels. *J. Appl. Physiol.*, 16:64, 1961.

Pietra, G. G., *et al.* Bronchial veins and pulmonary edema. *In* Fishman, A. P., and Renkin, E. M. (eds.) Pulmonary Edema. Baltimore, Waverly Press, 1979, p. 195.

Reed, J. H., Jr., and Wood, E. H. Effect of body position on vertical distribution of pulmonary blood flow. *J. Appl. Physiol.*, 29:303, 1970.

Reeves, J. T. Pulmonary vascular response to high altitude residence. *Cardiovasc. Clin.*, 5:81, 1973.

Staub, N. C. Pulmonary edema. Physiol. Rev., 54:678, 1974.

Staub, N. C. Pulmonary edema due to increased microvascular permeability. *Annu. Rev. Med.*, 32:291, 1981.

Staub, N. C. (ed.) Lung Water and Solute Exchange. New York: Marcel Dekker, 1978.

Staub, N. C. Pathways for fluid and solute fluxes in pulmonary edema. *In* Fishman, A. P., and Renkin, E. M. (eds.) Pulmonary Edema. Baltimore, Waverly Press, 1979, p. 113.

Taylor, A. E., *et al.* Permeability of the alveolar membrane to solutes. *Circ. Res.*, 16:353, 1965.

Taylor, A. E., et al. Na^{24} space, D_2O space, and blood volume in isolated dog lung. *Am. J. Physiol.*, 211:66, 1966.

West, J. B. Ventilation/Blood Flow and Gas Exchange. Oxford, Blackwell Scientific Publications, 1965.

West, J. B. Blood flow to the lung and gas exchange. *Anesthesiology*, 41:124, 1974.

Widimsky, J. (ed.) Pulmonary Hypertension. New York: S. Karger, 1975.

25

The Coronary Circulation and Ischemic Heart Disease

Approximately one third of all deaths in the affluent society of the Western world result from coronary artery disease, and almost all elderly persons have at least some impairment of the coronary artery circulation. For this reason, the normal and pathological physiology of the coronary circulation is one of the most important subjects in the entire field of medicine. The purpose of this chapter is to present this subject, emphasizing also the physiology of coronary occlusion and myocardial infarction.

NORMAL CORONARY BLOOD FLOW AND ITS VARIATIONS

PHYSIOLOGIC ANATOMY OF THE CORONARY BLOOD SUPPLY

Figure 25–1 illustrates the heart with its coronary blood supply Note that the main coronary arteries lie on the surface of the heart, and small arteries penetrate into the cardiac muscle mass It is almost entirely through these arteries that the heart receives its nutritive blood supply. Only the inner 75 to 100 microns of the endocardial surface can obtain significant amounts of nutrition directly from the blood in the cardiac chambers.

The *left coronary artery* supplies mainly the anterior and lateral portions of the left ventricle, whereas the *right coronary artery* supplies most of the right ventricle as well as the posterior part of the left ventricle in 80 to 90 per cent of all persons. In about half of all human beings, more blood flows through the right coronary artery than through the left, in about 30 per cent the arteries are about equal, and in 20 per cent the left artery predominates

Most of the venous blood flow from the left ventricle leaves by way of the *coronary sinus*—which is about 75 per cent of the total coronary blood flow—and most of the venous blood from the right ventricle flows through the small *anterior cardiac veins* directly into the right atrium and are not connected with the coronary sinus A small amount of coronary blood flows back into the heart through *thebesian veins*, which empty directly into all chambers of the heart.

NORMAL CORONARY BLOOD FLOW

The resting coronary blood flow in the human being averages approximately 225 ml per minute, which is about 0.7 to 0.8 ml per gram of heart muscle, or 4 to 5 per cent of the total cardiac output.

In strenuous exercise the heart increases its cardiac output as much as four- to sixfold, and it pumps this blood against a higher than normal arterial pressure. Consequently, the work output of the heart under severe conditions may increase as much as six- to eightfold. The coronary blood flow increases three- to fourfold to supply the extra nutrients needed by the heart. Obviously, this increase is not as much as the increase in work load, which means that the ratio of coronary blood flow to energy expenditure by the heart decreases. However, the "efficiency" of cardiac utilization of energy increases to make up for this relative deficiency of blood supply.

Phasic Changes in Coronary Blood Flow—Effect of Cardiac Muscle Compression. Figure 25–2 illustrates the average blood flow *through the small nutrient vessels* of the left ventricular coronary system in milliliters per minute in the human heart during systole and diastole as *calculated* from experiments in lower animals. Note from this diagram that the blood flow in the left ventricle falls to a low value during *systole*, which is opposite to the flow in all other vascular beds of the body. The reason for this is the strong compression of the left ventricular muscle around the intramuscular vessels during systole.

During *diastole*, the cardiac muscle relaxes completely and no longer obstructs the blood flow through the left ventricular capillaries, so that blood now flows rapidly during all of diastole.

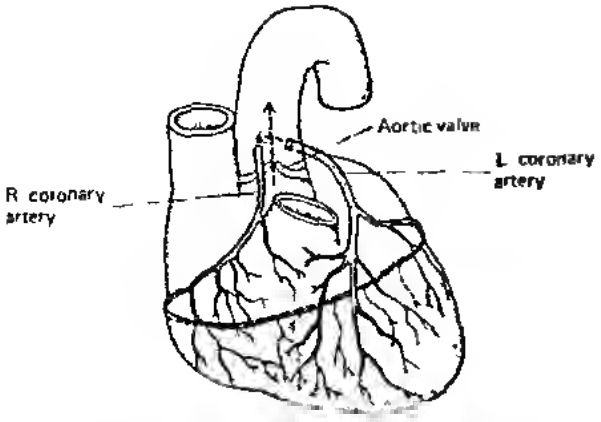


Figure 25–1. The coronary vessels

Blood flow through the coronary capillaries of the right ventricle also undergoes phasic changes during the cardiac cycle, but, because the force of contraction of the right ventricle is far less than that of the left ventricle, the inverse phasic changes are only partial in contrast to those in the left ventricle

Epicardial Versus Subendocardial Blood Flow—Effect of Intramyocardial Pressure. During cardiac contraction all the cardiac muscle squeezes toward the centers of the ventricles That is, the ventricular muscle adjacent to the ventricular chambers (the subendocardial muscle) squeezes the blood in the ventricle, the muscle in the middle layer of the ventricle squeezes the blood in the ventricle *and also the subendocardial muscle*, and the outermost muscle squeezes both the middle and the subendocardial muscle as well as the blood in the ventricle Therefore, during systole, a tissue pressure gradient develops within the heart muscle itself, with the pressure in the subendocardial muscle almost as great as the pressure inside the ventricle, while the pressure in the outer layer of the heart muscle is only slightly greater than atmospheric pressure The importance of this pressure gradient is that *the intramyocardial pressure in the inner layers of the heart muscle is so much greater than in the outer layers that it compresses the subendocardial blood vessels far more than the outer vessels*

Figure 25–3 illustrates the special arrangement of the coronary vessels at different depths in the heart, showing on the surface of the cardiac muscle the large epicardial coronary arteries that supply the heart Smaller intramuscular arteries penetrate the muscle, supplying the needed nutrients en route to the endocardium Then lying immediately beneath the endocardium is a plexus of subendocardial arteries During systole, blood flow through the subendocardial plexus of the left ventricle, where the contractile force of the muscle is very great, falls almost to zero To compensate for this almost total lack of flow during systole, the subendocardial arterial plexus is more extensive than the nutrient arteries in the middle and outer layers of the heart Therefore, during diastole, blood flow in the subendocardial arteries is considerably greater than is blood flow in the outermost arteries Later in the chapter we will see that this peculiar difference between blood flow in the epicardial and subendocardial arteries plays an important role in certain types of coronary ischemia

CONTROL OF CORONARY BLOOD FLOW

LOCAL METABOLISM AS THE PRIMARY CONTROLLER OF CORONARY FLOW

Blood flow through the coronary system is regulated almost entirely by vascular response to the local needs of the cardiac musculature for nutrition. This mechanism works equally well when the nerves to the heart are intact or are removed That is, whenever the vigor of contraction is increased, regardless of cause, the rate of coronary blood flow simultaneously increases, and, conversely, decreased activity is accompanied by decreased coronary flow It is immediately obvious that this local regulation of blood flow is almost identical with that occurring in many other tissues, especially in the skeletal muscles of all the body.

Oxygen Demand as a Major Factor in Local Blood Flow Regulation. Blood flow in the coronaries is regulated almost exactly in proportion to the need of the cardiac musculature for oxygen. Even in the normal resting state, 65 to 70 per cent of the oxygen in the arterial blood is removed as the blood passes through the heart; and, because not much oxygen is left, not much additional oxygen can be supplied to the heart musculature unless the blood flow increases. Fortunately, the blood flow does increase, and almost directly in proportion to the metabolic consumption of oxygen by the heart

Yet, the exact means by which increased oxygen consumption causes coronary dilatation has not been determined It is speculated that a decrease in the oxygen concentration in the heart causes vasodilator substances to be released from the muscle cells, and that these dilate the arterioles The substance with the greatest vasodilator propensity is *adenosine* In the presence of very low

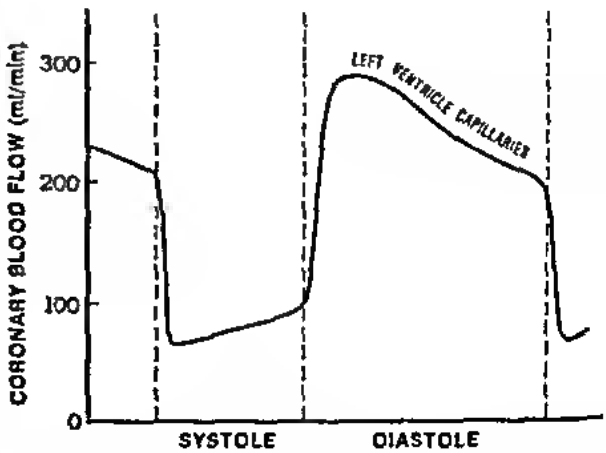


Figure 25–2. Phasic flow of blood through the coronary capillaries of the human left ventricle (as extrapolated from studies in dogs)

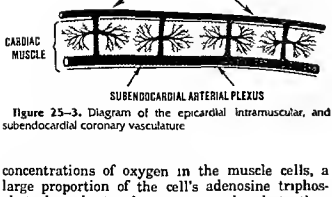


Figure 25–3. Diagram of the epicardial intramuscular, and subendocardial coronary vasculature

concentrations of oxygen in the muscle cells, a large proportion of the cell's adenosine triphosphate degrades to adenosine monophosphate; then small portions of this are further degraded to release adenosine into the tissue fluids of the heart muscle. After the adenosine causes vasodilatation it is destroyed within a few seconds so that it cannot go elsewhere in the circulation to cause unwarranted vasodilatation.

Adenosine is not the only vasodilator product that has been identified. Others include adenosine phosphate compounds, potassium ions, hydrogen ions, carbon dioxide, bradykinin, and possibly prostaglandins.

Yet, difficulties still exist with the vasodilator hypothesis. For one, agents that block or partially block the vasodilator effect of adenosine do not prevent the coronary vasodilatation in response to increased muscle activity. Second, studies in skeletal muscle have shown that adenosine can maintain vascular dilatation for only one to three hours, and yet muscle activity will still dilate the local blood vessels even when the adenosine can no longer dilate them.

Therefore, yet another theory to explain the coronary artery dilatation should be remembered until proved false: in the absence of adequate amounts of oxygen in the cardiac muscle, not only does the muscle itself suffer oxygen deficiency but so do the arteriolar muscle walls as well This could easily cause local vasodilatation because of lack of the required energy to keep the coronary vessels contracted against the high arterial pressure. But this, too, has its problems, because the coronary arteries require only minute amounts of oxygen to maintain full contraction.

The Determinants of Oxygen Consumption. Since the rate of oxygen consumption is the major factor that determines coronary blood flow, it is important to know the different factors that can alter myocardial oxygen consumption.

In general, the rate of cardiac oxygen consumption is closely related to work performed by the heart—the greater the work, the greater the oxygen consumption, and consequently the greater the coronary flow. However, this is not exactly true for the following reasons: Work output of the heart is determined by the arterial pressure against which the heart is pumping times the cardiac output pumped per minute. When the pressure increases, the oxygen consumption increases almost directly in proportion to the increase in pressure. On the other hand, when the cardiac output increases without an increase in pressure, the oxygen consumption increases only a slight to moderate amount rather than in direct proportion to the increase in cardiac output.

Therefore, an important principle of heart function is that myocardial energy consumption and oxygen usage are very nearly proportional to *tension* times *the time that the tension is maintained during the cardiac cycle.* One of the important consequences of this principle is the following: When the heart becomes dilated because of heart disease, the heart muscle must then generate greatly increased tension to pump against even the normal arterial pressure. This results from the *law of Laplace*, which states that the tension required to generate a given pressure increases in proportion to the diameter of the heart This necessity for increased tension means that, despite the diseased condition of this heart, the heart must expend greatly increased energy and use extra oxygen to provide the required pumping, which further exacerbates the heart's debility.

Other Causes of Increased Oxygen Consumption. Other factors that increase cardiac oxygen consumption are stimulation of the heart by epinephrine, norepinephrine, thyroxine, digitalis, and calcium ions and increased temperature of the heart. All these factors increase the metabolic activity of the cardiac muscle fibers themselves, which in turn increases the rate of oxygen usage even though they might not increase the work output of the heart. They also increase coronary blood flow approximately in proportion to the increase in rate of oxygen usage

Importance of the Increase in Coronary Blood Flow in Response to Myocardial Oxygen Usage. The resting heart extracts most of the oxygen from the coronary blood as it flows through the heart muscle, and very little of the heart's oxygen need can be met by additional extraction of oxygen from the coronary blood. Therefore, the only significant way in which the heart can be supplied with additional amounts of oxygen is through an increase in blood flow. Consequently, it is essential that the coronary blood flow increase whenever the heart muscle demands additional oxygen. When the coronary blood flow fails to increase appropriately, the strength of the muscle diminishes rapidly and drastically, often causing acute heart failure. Also, this *relative* ischemia of the muscle can cause severe pain, called *angina pain*, which will be discussed later in the chapter.

Reactive Hyperemia in the Coronary System. If the coronary flow to the heart is completely occluded for a few seconds to a few minutes and then suddenly unoccluded, the blood flow increases to as high as 3 to 4 times normal, as shown in Figure 25–4. It remains high for a few seconds to several minutes, depending on the

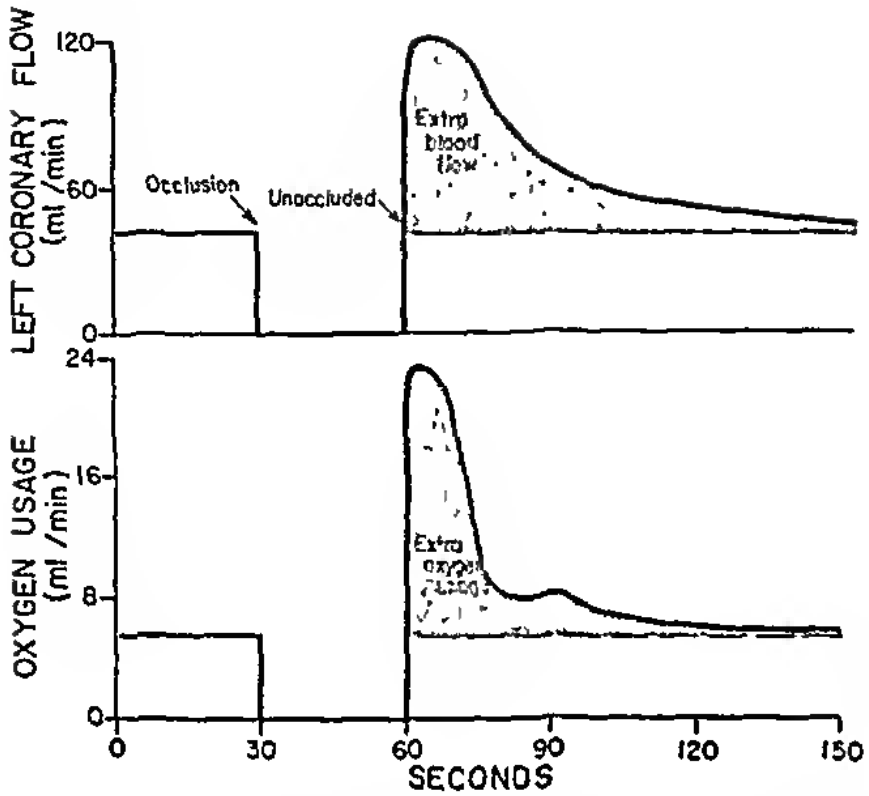


Figure 25–4. Reactive hyperemia in th coronary system caused by a 30 secon period of coronary occlusion. Note the extr blood flow, called "reactive hyperemia," an the extra oxygen usage after the period c occlusion was over

period of occlusion This extra flow of blood is called reactive hyperemia, which was explained in Chapter 20 During the period of excess flow, the heart removes a *large amount of extra oxygen from the blood, as shown* by the bottom curve of the figure, to make up for the deficiency of oxygen during the period of occlusion Reactive hyperemia is simply another manifestation of the ability of the coronary system to adjust its flow to the metabolic needs of the heart

NERVOUS CONTROL OF CORONARY BLOOD FLOW

Stimulation of the autonomic nerves to the heart can affect coronary blood flow in two ways—directly and indirectly The direct effects result from direct action of the nervous transmitter substances, acetylcholine and norepinephrine, on the coronary vessels themselves The indirect effects result from secondary changes in coronary blood flow caused by increased or decreased activity of the heart

The indirect effects play by far the more important role in normal control of coronary blood flow. Thus, sympathetic stimulation, which releases norepinephrine, increases both heart rate and heart contractility as well as its rate of metabolism. In turn, the increased activity of the heart sets off local blood flow regulatory mechanisms for dilating the coronary vessels, the blood flow increasing approximately in proportion to the metabolic needs of the heart muscle. In contrast, vagal stimulation, with its release of acetylcholine, slows the heart and also has a slight depressive effect on cardiac contractility. Both these effects decrease cardiac oxygen consumption and therefore indirectly constrict the coronaries.

Direct Effects of Nervous Stimuli on the Coronary Vasculature. The distribution of parasympathetic (vagal) nerve fibers to the ventricular coronary system is so slight that parasympathetic stimulation has an almost negligible direct effect on coronary blood flow though a very slight effect on dilating the coronaries

On the other hand, there is much more extensive sympathetic innervation of the coronary vessels. In Chapter 57 we shall see that the sympathetic transmitter substances norepinephrine and epinephrine can have either dilator or constrictor effects, depending on the presence or absence of specific receptors in the blood vessel walls The constrictor receptors are called *alpha receptors*, and the dilator receptors are called *beta receptors* Both alpha and beta receptors are known to exist in the coronary vessels. In general, the epicardial coronary vessels have a preponderance of alpha receptors whereas the intramuscular arteries have a preponderance of beta receptors Therefore sympathetic stimulation can cause either slight overall coronary constriction or dilatation, but usually a little more constriction Yet, in some persons, the alpha vasoconstrictor effects seem to be disproportionately severe, and these persons can have vasospastic myocardial ischemia during periods of excess sympathetic drive, often with resultant anginal pain

It must be pointed out again, however, that metabolic factors—especially myocardial oxygen consumption—are the major controllers of myocardial blood flow. Therefore, whenever nervous stimulation alters the coronary blood flow, the metabolic factors usually return the flow most of the way back toward normal within *seconds or minutes.*

THE SUBSTRATES OF CARDIAC METABOLISM

The general principles of cellular metabolism, which will be discussed in Chapters 67 through 70, apply to cardiac muscle, though there are some quantitative differences. Most importantly, under resting conditions cardiac muscle normally utilizes mainly fatty acids for its energy, with approximately 70 per cent of the normal metabolism being derived from these. However, as is true of other tissues, under anaerobic or ischemic conditions cardiac metabolism must call upon the anaerobic glycolysis mechanism for energy. This, unfortunately, can supply very little extra energy in relation to the large energy requirements of the heart. Also, it utilizes tremendous quantities of the blood glucose and at the same time forms large amounts of lactic acid in the cardiac tissue, which is probably one of the causes of cardiac pain in cardiac ischemic conditions, as will be discussed later in the chapter.

As is true in other tissues, more than 95 per cent of the metabolic energy liberated from the foods is used to form ATP in the mitochondria. This ATP in turn supplies the energy for cellular function. In coronary ischemia, the ATP degrades to ADP, AMP, and adenosine. Because the cell membrane is permeable to adenosine, some of this is lost from the muscle cells into the circulating blood. This released adenosine is believed to be one of the substances that causes dilatation of the coronary arterioles during coronary hypoxia, as discussed earlier. However, the loss of adenosine also has a very serious cellular consequence. Within as little as half an hour of severe coronary ischemia, as occurs after a myocardial infarct or during cardiac arrest, about half of the adenine base can be lost from the cardiac muscle cells. Furthermore, this can be replaced by new synthesis of adenosine at a rate of only 2 per cent per hour. Therefore, once a serious bout of ischemia has persisted for half an hour or so, relief of the coronary ischemia may be too late to save the lives of the cardiac cells. This is almost certainly the major cause of cardiac cellular death following myocardial ischemia and also one of the most important causes of cardiac debility in the late stages of circulatory shock, as will be discussed in Chapter 28.

ISCHEMIC HEART DISEASE

The single most common cause of death in Western culture is ischemic heart disease, which results from insufficient coronary blood flow. Approximately 35 per cent of all human beings in the United States die of this cause. Some deaths occur suddenly as a result of an acute coronary occlusion or of fibrillation of the heart, whereas others occur slowly over a period of weeks to years as a result of progressive weakening of the heart pumping process. In the present chapter we will discuss the coronary ischemia problem itself, as well as acute coronary occlusion and myocardial infarction. In the following chapter we will discuss congestive heart failure, the most frequent cause of which is progressive coronary ischemia.

Atherosclerosis as the Cause of Ischemic Heart Disease. The most frequent cause of diminished coronary blood flow is atherosclerosis. Although the atherosclerotic process will be discussed in connection with lipid metabolism in Chapter 68, briefly, this process is the following: In certain persons who have a genetic predisposition to atherosclerosis or in persons who eat *excessive quantities of cholesterol* and *other fats*, large quantities of cholesterol gradually become deposited beneath the intima at many points in the arteries. Later, these areas of deposit become invaded by fibrous tissue, and they also frequently become calcified. The net result is the development of atherosclerotic plaques that protrude into the vessels and either block or partially block blood flow.

A very common site for development of atherosclerotic plaques is the first few centimeters of the coronary arteries.

Acute Coronary Occlusion. Acute occlusion of a coronary artery frequently occurs in a person who already has serious underlying atherosclerotic coronary heart disease, but almost never in a person with a normal coronary circulation. This condition can result from any one of several different effects, two of which are the following:

1. The atherosclerotic plaque can cause a local blood clot called a *thrombus*, which in turn occludes the artery. The thrombus usually begins where the plaque has grown so much that it has broken through the intima, thus coming in contact with the flowing blood. Because the plaque presents an unsmooth surface to the blood, platelets begin to adhere to it, fibrin begins to be deposited, and blood cells become entrapped and form a clot that grows until it occludes the vessel. Or occasionally the clot breaks away from its attachment on the atherosclerotic plaque and flows to a more peripheral branch of the coronary arterial tree where it blocks the artery at that point. A thrombus that flows along the artery in this way and occludes the vessel more distally is called an *embolus*.
2. It is believed by many clinicians that local spasm of a coronary artery can also cause sudden occlusion. The spasm might result from irritation of the smooth muscle of the arterial wall by the edges of an atherosclerotic plaque, or it might result from nervous reflexes that cause the coronary contraction. The spasm may then lead to *secondary* thrombosis of the vessel.

Collateral Circulation in the Heart. The degree of damage to the heart caused either by slowly developing atherosclerotic constriction of the coronary arteries or by sudden occlusion is determined to a great extent by the degree of collateral circulation that is already developed or that can develop within a short period of time *after the occlusion*.

Unfortunately, in a normal heart, almost no communications exist among the larger coronary arteries. But many anastomoses do exist among the smaller arteries sized 20 to 250 microns in diameter, as shown in Figure 25-5.

When a sudden occlusion occurs in one of the larger coronary arteries, these small anastomoses dilate within a few seconds. The blood flow through these minute collaterals is usually less than one half that needed to keep alive the cardiac muscle that they supply; and unfortunately the diameters of the collateral vessels do not enlarge farther for the next 8 to 24 hours. But then collateral flow does begin to increase, doubling by the second or third day and often reaching normal or almost normal coronary flow in the previously ischemia muscle within about one month; it is capable of increasing even further with increased metabolic loads. It is because of these developing collateral channels that a

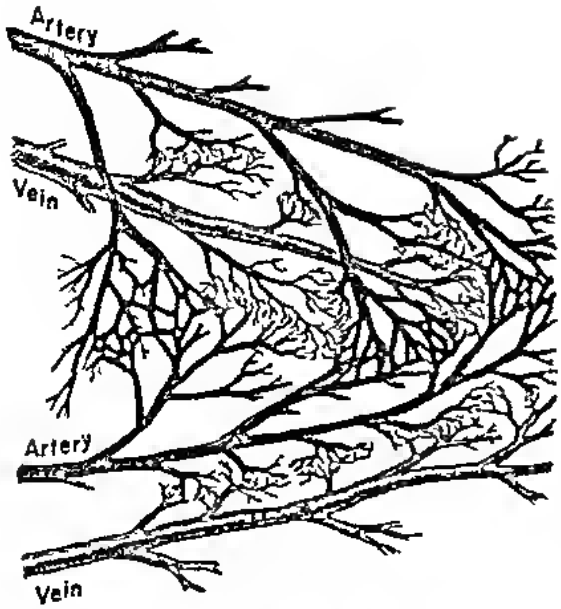


Figure 25–5. Minute anastomoses of the coronary arterial system

recovers from various types of coronary occlusion when the area of muscle involved is not too great.

When atherosclerosis constricts the coronary arteries slowly over a period of many years rather than suddenly, collateral vessels can develop at the same time that the atherosclerosis does. Therefore, the person may never experience an acute episode of cardiac dysfunction. Eventually, however, the sclerotic process develops beyond the limits of even the collateral blood supply to provide the needed blood flow, and even the collaterals sometimes develop atherosclerosis. When this occurs, the heart muscle becomes severely limited in its work output, often so much so that the heart cannot pump even the normally required amounts of blood flow. This is the most common cause of cardiac failure and occurs in vast numbers of older persons.

MYOCARDIAL INFARCTION

Immediately after an acute coronary occlusion, blood flow ceases in the coronary vessels beyond the occlusion except for small amounts of collateral flow from surrounding vessels. The area of muscle that has either zero flow or so little flow that it cannot sustain cardiac muscle function is said to be *infarcted*. The overall process is called a *myocardial infarction*.

Soon after the onset of the infarction small amounts of collateral blood seep into the infarcted area, and this, combined with progressive dilatation of the local blood vessels, causes the area to become overfilled with stagnant blood. Simultaneously, the muscle fibers utilize the last vestiges of the oxygen in the blood, causing the hemoglobin to become totally reduced and very dark blue in color. Therefore, the infarcted area takes on a dark bluish-brown hue, and the blood vessels of the area appear to be engorged despite the lack of blood flow. In later stages, the vessels become highly permeable and [illegible] fluid, the tissue becomes edematous, and the cardiac [illegible] cells begin to swell because of diminished cellu- metabolism. Within a few hours of almost no blood supply the cells die.

Cardiac muscle requires approximately 1.3 ml of oxygen per 100 grams of muscle tissue per minute simply to remain alive. This is in comparison with approximately 8 ml of oxygen per 100 grams delivered to the normal resting left ventricle each minute. Therefore, if there is even as much as 15 to 30 per cent of normal resting coronary blood flow, the muscle will not die. In the central portion of a large infarct, however, the blood flow is usually less than this so that this muscle does die.

Subendocardial Myocardial Infarction. Myocardial infarction frequently occurs in the subendocardial muscle even when the epicardial portions of the heart muscle remain uninfarcted. This form of infarction occurs especially when the diastolic arterial pressure is very low or when the diastolic intraventricular pressure is very high. From the discussion earlier in the chapter of the subendocardial arterial plexus, it was pointed out that most blood flow into this area of the heart occurs during diastole. Therefore, when the diastolic arterial pressure is very low—as occurs in patients who have aortic regurgitation, patent ductus arteriosus, or to a lesser extent arteriosclerosis—one can expect a high incidence of subendocardial myocardial infarction.

CAUSES OF DEATH FOLLOWING ACUTE CORONARY OCCLUSION

The four major causes of death following acute myocardial infarction are decreased cardiac output; damming of blood in the pulmonary or systemic veins with death resulting from edema, especially pulmonary edema, fibrillation of the heart, and, occasionally, rupture of the heart.

Decreased Cardiac Output—Cardiac Shock. When some of the cardiac muscle fibers are not functioning at all and others are too weak to contract with great force, the overall pumping ability of the affected ventricle is proportionately depressed. Indeed, the overall pumping strength of the heart is often decreased more than one might expect because of the phenomenon of *systolic stretch*, which is illustrated in Figure 25–6. When the normal portions of the ventricular muscle contract, the ischemic muscle, whether this be dead or simply nonfunctional, instead of contracting is actually forced outward by the pressure that develops inside the ventricle. Therefore, much of the pumping force of the ventricle is dissipated by bulging of the area of nonfunctional cardiac muscle.

When the heart becomes incapable of contracting with sufficient force to pump enough blood into the arterial tree, cardiac failure and death of the peripheral tissues ensue as a result of peripheral ischemia. This condition is called *coronary shock, cardiogenic shock, cardiac shock*, or *low cardiac output failure*. It will be discussed in the following chapter. Cardiac shock almost always occurs when more than 40 per cent of the left ventricle is infarcted. Death occurs in about 85 per cent of those patients who develop cardiac shock.

Damming of Blood in the Venous System. When the heart is not pumping blood forward, it must be damming blood in the blood vessels of the lungs or in the systemic circulation. Obviously, this increases both the left and right atrial pressures and also leads to increased capillary pressures, particularly in the lungs. These effects often cause very little difficulty during the first few hours after the myocardial infarction. Instead,

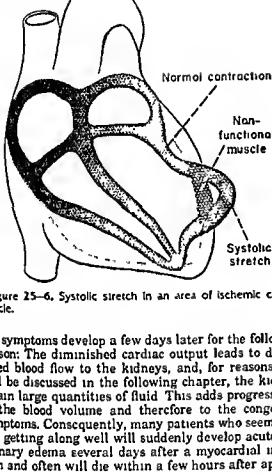


Figure 25–6. Systolic stretch in an area of ischemic cardiac muscle.

the symptoms develop a few days later for the following reason: The diminished cardiac output leads to diminished blood flow to the kidneys, and, for reasons that will be discussed in the following chapter, the kidneys retain large quantities of fluid This adds progressively to the blood volume and therefore to the congestive symptoms. Consequently, many patients who seemingly are getting along well will suddenly develop acute pulmonary edema several days after a myocardial infarction and often will die within a few hours after appearance of the initial edema symptoms

Rupture of the Infarcted Area. During the first day of an acute infarct there is little danger of rupture of the ischemic portion of the heart, but a few days after a large infarct occurs, the dead muscle fibers begin to degenerate, and the dead heart musculature is likely to become very thin. If this happens, the dead muscle bulges outward severely with each heart contraction, an effect called systolic stretch as explained above, and this stretch becomes greater and greater until finally the heart ruptures. In fact, one of the means used in assessing the progress of a severe myocardial infarction is to record by radiograph whether the degree of systolic stretch is getting worse

When a ventricle does rupture, the loss of blood into the pericardial space causes rapid development of *cardiac tamponade*—that is, compression of the heart from the outside by blood collecting in the pericardial cavity. Because the heart is compressed, blood cannot flow into the right atrium, and the patient dies of suddenly decreased cardiac output

Fibrillation of the Ventricles Following Myocardial Infarction. Many persons who die of coronary occlusion die because of ventricular fibrillation. The tendency to develop fibrillation is especially great following a large infarction, but fibrillation can occur following small occlusions as well. Indeed, many patients with coronary *insufficiency die acutely from fibrillation* without any infarction at all.

There are two especially dangerous periods during which fibrillation is most likely to occur. The first is during the first ten minutes after the infarction occurs. Then there is a short period of relative safety, followed by a secondary period of cardiac irritability beginning an hour or so later and lasting for another few hours However, fibrillation can occur for many days after the infarct

At least four different factors enter into the tendency for the heart to fibrillate·

First, acute loss of blood supply to the cardiac muscle causes rapid depletion of potassium from the ischemic musculature This increases the potassium concentration in the extracellular fluids surrounding the cardiac muscle fibers Experiments in which potassium has been injected into the coronary system have demonstrated that an elevated extracellular potassium concentration increases the irritability of the cardiac musculature.

Second, ischemia of the muscle causes an "injury current," which was described in Chapter 16 in relation to electrocardiograms in patients with acute myocardial infarction. The ischemic musculature cannot repolarize its membranes so that this muscle remains negative with respect to the normal polarized cardiac muscle membrane Therefore, electrical current flows from this ischemic area of the heart to the normal area and can elicit abnormal impulses which can cause fibrillation.

Third, powerful sympathetic reflexes develop following massive infarction, principally because the heart does not pump an adequate volume of blood into the arterial tree The sympathetic stimulation also increases the irritability of the cardiac muscle and thereby predisposes to fibrillation.

Fourth, the myocardial infarction itself often causes the ventricle to dilate excessively This increases the pathway length for impulse conduction in the heart and also frequently causes abnormal conduction pathways around the infarcted area of the cardiac muscle Both of these effects predispose to development of circus movements. As was discussed in Chapter 14, any prolongation of the conduction pathway in the ventricles will allow an impulse to re-enter muscle that is already recovering from refractoriness, thereby initiating a subsequent cycle of excitation and causing the process to continue on and on

THE STAGES OF RECOVERY FROM ACUTE MYOCARDIAL INFARCTION

The upper part of Figure 25–7 illustrates the effects of acute coronary occlusion, on the left, in a patient with a small area of muscle ischemia and, on the right, in a patient with a large area of ischemia. When the area of ischemia is small, little or no death of the muscle cells

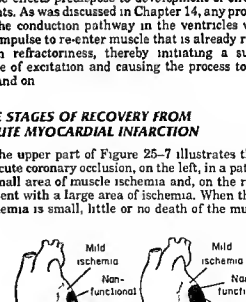


Figure 25–7. *Top.* Small and large areas of coronary ischemia. *Bottom* Stages of recovery from myocardial infarction.

may occur, but part of the muscle often does become temporarily nonfunctional because of inadequate nutrition to support muscle contraction.

When the area of ischemia is large, some of the muscle fibers in the very center of the area die rapidly, within one to three hours in an area of total cessation of coronary blood supply Immediately around the dead area is a nonfunctional area because of failure of contraction and usually also failure of impulse conduction. Then, extending circumferentially around the nonfunctional area is an area that is still contracting but weakly so because of mild ischemia

Replacement of Dead Muscle by Scar Tissue. In the lower part of Figure 25–7, the various stages of recovery following a myocardial infarction are illustrated Shortly after the occlusion, the muscle fibers die in the very center of the ischemic area Then during the ensuing days, this area of dead fibers grows because many of the marginal fibers finally succumb to the prolonged ischemia At the same time, owing to the enlargement of the collateral arterial channels growing into the outer rim of the infarcted area, the nonfunctional area of muscle becomes smaller and smaller After a few days to three weeks most of the nonfunctional area of muscle becomes functional again or dead—one or the other In the meantime, fibrous tissue begins developing among the dead fibers, for ischemia stimulates growth of fibroblasts and promotes development of greater than normal quantities of fibrous tissue. Therefore, the dead muscular tissue is gradually replaced by fibrous tissue Then, because it is a general property of fibrous tissue to undergo progressive elastomeric contraction and dissolution, the size of the fibrous scar may become smaller over a period of several months to a year

Finally, the normal areas of the heart gradually hypertrophy to compensate at least partially for the lost cardiac musculature By these means the heart recovers

Value of Rest in Treating Myocardial Infarction. The degree of cellular death is determined by the *degree of ischemia × the degree of metabolism* of the heart muscle When the metabolism of the heart muscle is greatly increased, such as during exercise, in severe emotional strain, or as a result of fatigue, the heart musculature needs increased oxygen and other nutrients for sustaining its life Furthermore, anastomotic blood vessels that supply blood to ischemic areas of the heart must also still supply the areas of the heart that they normally supply When the heart becomes excessively active, the vessels of the normal musculature become greatly dilated, and this allows most of the blood flowing into the coronary vessels to flow through the normal muscle tissue, thus leaving little blood to flow through the small anastomotic channels into the ischemic area, so that the ischemic condition is worsened This condition is called the "coronary steal" syndrome. Consequently, one of the most important factors in the treatment of a patient with myocardial infarction is observance of absolute rest The greater the degree of rest during the first few days or weeks following sudden infarction, the less the degree of cellular death

OF THE HEART

MYOCARDIAL INFARCTION

Occasionally, a heart that has recovered from a large myocardial infarction returns to full functional capability, but more frequently its pumping capability is permanently decreased below that of a normal heart. This does not mean that the person is necessarily a cardiac invalid or that the resting cardiac output is depressed below normal, because the normal person's heart is capable of pumping 400 per cent more blood per minute than the body requires—that is, a person has a "cardiac reserve" of 400 per cent. Even when the cardiac reserve is reduced to as little as 100 per cent, the person can still perform normal activity of a quiet, restful type, but not strenuous exercise that would overload the heart.

The effect of myocardial infarction and other myocardial debilities on cardiac output is discussed in the following chapter in relation to the regulation of cardiac output in cardiac failure

PAIN IN CORONARY DISEASE

Normally, a person cannot "feel" his or her heart, but ischemic cardiac muscle does exhibit pain sensation. Exactly what causes this pain is not known, but it is believed that ischemia causes the muscle to release acidic substances such as lactic acid or other pain-promoting products such as histamine or kinins that are not removed rapidly enough by the slowly moving blood. The high concentrations of these abnormal products then stimulate the pain endings in the cardiac muscle, and pain impulses are conducted through the sympathetic afferent nerve fibers into the central nervous system.

ANGINA PECTORIS

In most persons who develop progressive constriction of their coronary arteries, cardiac pain, called *angina pectoris*, begins to appear whenever the load on the heart becomes too great in relation to the coronary blood flow This pain is usually felt beneath the upper sternum and is often also referred to surface areas of the body, most often to the left arm and left shoulder but also frequently to the neck and even to the side of the face or to the opposite arm and shoulder. The reason for this distribution of pain is that the heart originates during embryonic life in the neck, as do the arms Therefore, both of these structures receive pain nerve fibers from the same spinal cord segments.

In general, most persons who have chronic angina pectoris feel the pain when they exercise, also when they experience emotions that increase metabolism of the heart or temporarily constrict the coronary vessels because of sympathetic vasoconstrictor nerve signals. Usually the pain lasts for only a few minutes However, some patients have such severe and lasting ischemia that the pain is present all the time The pain is frequently described as dull, pressing, and constricting; it is of such quality that it usually makes the patient stop all activity and come to a complete state of rest

When a person has frequent attacks of angina, the likelihood of developing an acute coronary occlusion is generally very great.

Treatment with Vasodilator Drugs. Several vasodilator drugs, when administered during an acute anginal attack, will often give immediate relief from the pain Two commonly used drugs are nitroglycerin and amyl nitrite. The original theory for use of these drugs was that they caused coronary vasodilatation and that this relieved the pain by increasing coronary blood flow. However, in the normal person the drugs have almost

no effect on dilatation of the coronary *arterioles* and da not cause a measurable increase in the coronary blood flow. But they do have two other effects that probably explain their efficacy in relieving anginal pain. First, they cause marked venous dilatation throughout the body, which reduces venous return to the heart, cardiac output, and arterial pressure. Therefore, the work load of the heart is immediately diminished, thereby decreasing the formation of the ischemic products that are presumed to be the cause of the pain. Second, even though these drugs do not dilate the coronary arterioles, they do dilate the large *epicardial* coronary arteries In the person who has severely constricted epicardial arteries, this slight dilatation can probably be of additional benefit in relieving the anginal pain even though in the normal heart such dilatation causes no measurable increase in coronary blood flow.

Treatment with Sympathetic Blocking Drugs. Another medical procedure for relieving anginal pain is to block the sympathetic nervous stimulation of the heart. This is most frequently accomplished by administering *propranolol*, which blocks the beta adrenergic receptors Stimulation of these receptors enhances cardiac activity. Therefore, when they are blocked, the level of cardiac muscle contractility and the heart rate both decrease, and the cardiac work output also decreases; at the same time, the heart's usage of oxygen diminishes as much as 10 to 20 per cent. In this way, the ratio of oxygen need to oxygen supply is often reduced enough to eliminate the pain

SURGICAL TREATMENT OF CORONARY DISEASE

Aortic-Coronary Bypass Surgery. In many patients with coronary ischemia, the constricted areas of the coronary vessels are located at only a few discrete points, and the coronary vessels beyond these points are of normal or almost normal size. A surgical procedure has been developed in the past 20 years, called *aortic-coronary bypass*, for anastomosing small vein grafts to the aorta and to the sides of the more peripheral coronary vessels. Usually, one to three such grafts are performed during the operation, each of which supplies a peripheral coronary artery beyond a block. However, on occasion, as many as five grafts have been inserted in a single person. The vein that is used for the graft is usually the long superficial saphenous vein removed from the leg of the recipient.

The acute results from this type of surgery have been especially good, causing this to be the most common cardiac operation performed. Anginal pain is relieved in most patients. Also, in patients whose hearts have not become too severely damaged prior to the operation, the coronary bypass procedure often can provide the patient with a normal survival expectation. On the other hand, if the heart has already been severely damaged, the bypass procedure is likely to be of little value.

Coronary Angioplasty. During the past ten years, a procedure has been employed to open partially blocked coronary vessels before they become totally occluded. This procedure, called *coronary artery angioplasty*, is the following: A very small catheter, about 1 millimeter in diameter, has a balloon surrounding it, beginning a few millimeters from its tip and extending backward along its shaft for 2 to 3 centimeters. The tip of this catheter is passed under radiographic guidance into the coronary system and pushed through the partially occluded artery until the balloon portion of the catheter straddles the partially occluded point. Then the balloon is inflated with several atmospheres of pressure, which stretches the diseased artery almost to the point of bursting. After this procedure is performed, the blood flow through the vessel usually increases as much as three to fourfold, and more than three quarters of the patients who undergo the procedure are relieved of the coronary ischemic symptoms for at least several years, though many of the patients still will require coronary bypass surgery eventually.

MEASUREMENT OF HUMAN CORONARY BLOOD FLOW

Obviously, it would be almost impossible to insert a flowmeter into the coronary circulatory system of human beings. However, a satisfactory indirect method for measuring coronary flow has been achieved, the principles of which are the following:

The blood flow through any organ of the body can be determined by means of the so-called Fick principle, which was discussed in Chapter 23 in relation to the measurement of cardiac output. That is, if a constituent of the blood is removed as it flows through an organ, then the blood flow can be determined from two measurements: first, the amount removed from each milliliter of blood as it goes through the organ and, second, the total quantity of the substance that is removed by the organ in a given period of time.

In measuring blood flow through the heart the subject suddenly begins to breathe nitrous oxide of a given concentration Samples of arterial blood are removed from any artery in the body and samples of venous blood are obtained from the coronary sinus through a catheter that has been passed into the sinus by way of the venous system The concentrations of the nitrous oxide in both of these bloods are then plotted for ten minutes, as shown in Figure 25–8. The quantity of nitrous oxide lost to the heart by each milliliter of blood as it passes through the heart at each instant of the experiment can be determined from the difference in height of the two curves. This is called the *A-V difference*. The A-V difference is determined each minute for a period of ten minutes and then averaged.

Now, to determine the amount of nitrous oxide that is removed from the blood by the cardiac muscle, one proceeds as follows: It will be noted from the curves of Figure 25–8 that the concentrations of nitrous oxide begin to reach an equilibrium level toward the end of the ten minute experiment. It is also known that the total amount of nitrous oxide absorbed by the cardiac muscle is directly proportional to its concentration in the arterial blood at the end of the ten minute run. A *proportionality factor* has been determined in experiments on isolated cardiac muscle so that at the end of ten minutes the total amount of nitrous oxide absorbed by each 100 grams of cardiac muscle can be determined by multiplying the arterial concentration of nitrous oxide times the proportionality factor.

Once the average *A-V difference* and the *amount of nitrous oxide absorbed* by each 100 grams of heart muscle have been determined, the coronary blood flow per 100 grams of muscle is then calculated by dividing the average A-V difference into the quantity absorbed

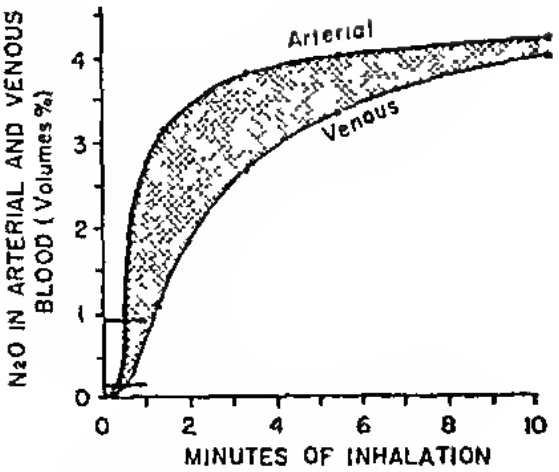


Figure 25–8. Method for measuring human coronary blood flow as explained in the text

per minute. As an example, if 8 ml of nitrous oxide is absorbed by 100 grams of heart in ten minutes and the average A-V difference during this period of time is 0.01 ml for each milliliter of blood, a total of 800 ml of blood would have passed through the 100 grams of heart during the ten minute interval. In other words, the blood flow would be 80 ml per 100 grams of heart per minute. Assuming a heart mass of 280 grams, the blood flow would be 224 ml per minute.

REFERENCES

Berne, R. M. and Rubio, R. Regulation of coronary blood flow. *Adv. Cardiol.* 12:303, 1974.

Berne, R. M., and Rubio, R. Coronary circulation. *In* Berne, R. M. *et al.* (eds.) Handbook of Physiology. Sec. 2, Vol. I. Baltimore: Williams & Wilkins, 1979, p. 873.

Bodem, G. and Dengler, H. J. (eds.) Cardiac Glycosides. New York: Springer-Verlag, 1978.

Braunwald, E. and Maroko, P. R. Limitation of Infarct Size. Chicago: Year Book Medical Publishers, 1978.

Budassi, S. A. (ed.) Cardiopulmonary Resuscitation. Germantown, Md.: Aspen Systems Corp., 1978.

Conti, C. R. Large vessel coronary vasospasm: Diagnosis, natural history and treatment. *Am. J. Cardiol.* 55:41B, 1985.

Epstein, S. E. *et al.* Dynamic coronary obstruction as a cause of angina pectoris: Implications regarding therapy. *Am. J. Cardiol.*, 55:61B, 1985.

Feigl, E. O. Coronary physiology. *Physiol. Rev.* 63:1, 1983.

Fletcher, G. F. and Cantwell, J. D. Exercise and Coronary Heart Disease: Role in Prevention, Diagnosis, Treatment. Springfield, Ill.: Charles C. Thomas, 1978.

Geddes, L. A. Cardiovascular Medical Devices. New York: John Wiley & Sons, 1984.

Gensini, G. G. (ed.) Coronary Arteriography. Mt. Kisco, N. Y.: Futura Publishing Co., 1985.

Gibbs, C. L. and Chapman, J. B. Cardiac energetics. *In* Berne, R. M., *et al.* (eds.) Handbook of Physiology. Sec. 2, Vol. I. Baltimore: Williams & Wilkins, 1979, p. 775.

Gregg, D. E. Coronary Circulation in Health and Disease. Philadelphia: Lea & Febiger, 1950.

Gregg, D. E. The natural history of coronary collateral development. *Circ. Res.* 35:335, 1974.

Guyton, R. A. and Daggett, W. M. The evolution of myocardial infarction: Physiological basis for clinical intervention. *Int. Rev. Physiol.*, 505, 1976.

Hall, J. I. (ed.) Differential Diagnosis of Chest Pain and Other Cardiac Symptoms. Mt. Kisco, N. Y.: Futura Publishing Co., 1983.

Hutchins, G. M. Pathological changes in aortocoronary bypass graft. *Annu. Rev. Med.*, 31:289, 1980.

Hutchinson: Prevention of Coronary Heart Disease. Chicago: Year Book Medical Publishers, 1985.

Iskandrian, A. S., and Hakki, A. H. Thallium-201 myocardial scintigraphy. *Am. Heart J.* 109:113, 1985.

Kaisner, S. (ed.) The Coronary Artery. New York, Oxford University Press, 1982.

Katz, A. M. Basic cellular mechanisms of action of the calcium-channel blockers. *Am. J. Cardiol.*, 55:2B, 1985.

Klocke, F. J., and Ellish, A. K. Control of coronary blood flow. *Annu. Rev. Med.* 31:489, 1980.

Lawrie: Coronary Artery Bypass Surgery. Chicago: Year Book Medical Publishers, 1985.

Leon, M. B., *et al.* Combination therapy with calcium-channel blockers and beta blockers for chronic stable angina pectoris. *Am. J. Cardiol.*, 55:69B, 1985.

Long, C. (ed.) Prevention and Rehabilitation in Ischemic Heart Disease. Baltimore: Williams & Wilkins, 1980.

Lüllmann, H. and Peters, T. Action of Cardiac Glycosides on the Excitation Contraction Coupling in Heart Muscle. New York: Fischer, 1979.

Manninen, V. and Halonen, P. I. (eds.) Sudden Coronary Death. New York: S. Karger, 1978.

McAllister, R. G. Jr., *et al.* Pharmacokinetics of calcium-entry blockers. *Am. J. Cardiol.*, 55:30B, 1985.

Neely, J. R. and Morgan, H. E. Relationship between carbohydrate and lipid metabolism and the energy balance of the heart. *Annu. Rev. Physiol.* 36:413, 1974.

Norris, R. M. Myocardial Infarction: Its Presentation, Pathogenesis and Treatment. New York: Churchill Livingstone, Inc., 1982.

Pajndas, G. *et al.* Coronary Angiography in the Medical and Surgical Treatment of Ischemic Heart Disease. New York: McGraw Hill, 1980.

Reimer, K. A. and Jennings, R. B. Effects of calcium-channel blockers on myocardial preservation during experimental acute myocardial infarction. *Am. J. Cardiol.* 55:107B, 1985.

Rigotti, N. A. *et al.* Exercise and coronary heart disease. *Annu. Rev. Med.* 34:391, 1983.

Roberts, R. and Marmor, A. T. Right ventricular infarction. *Annu. Rev. Med.* 34:377, 1983.

Rose, C. P. and Goresky, C. A. Interactions between capillary exchange, cellular entry, and metabolic sequestration processes in the heart. *In* Renkin, E. M. and Michel, C. C. (eds.) Handbook of Physiology. Sec. 2, Vol. IV. Bethesda: American Physiological Society, 1984, p. 761.

Roskamm, H., and Schumuziger, M. (eds.) Coronary Heart Surgery: A Rehabilitation Measure. New York: Springer Verlag, 1979.

Russell, R. O. Jr. (ed.) Hemodynamic Monitoring in a Coronary Intensive Care Unit. Mt. Kisco, N. Y.: Futura Publishing Co., 1981.

Sarnoff, S. J. *et al.* Hemodynamic determinants of oxygen consumption of the heart with special reference to the tension time index. *Am. J. Physiol.* 192:148, 1958.

Schaper, W. (ed.) The Pathophysiology of Myocardial Perfusion. New York, Elsevier/North Holland, 1979.

Shepherd, J. T. and Vanhoutte, P. M. Spasm of the coronary arteries: Causes and consequences (the scientist a viewpoint). *Mayo Clin. Proc.* 60:33, 1985.

Singh, B. N., *et al.* Second generation calcium antagonists: Search for greater selectivity and versatility. *Am. J. Cardiol.*, 55:214B, 1985.

Stein, E. A., *et al.* Coronary risk factors in the young. *Annu. Rev. Med.*, 32:601, 1981.

Stone, H. L. Control of the coronary circulation during exercise. *Annu. Rev. Physiol.* 45:213, 1983.

Stone, H. L., *et al.* Cardiac function after embolization of coronaries with microspheres. *Am. J. Physiol.* 204:16, 1963.

Stone, H. L., *et al.* Progressive changes in cardiovascular function after unilateral heart irradiation. *Am. J. Physiol.* 206:269, 1964.

Strauss, H. W. (ed.) Cardiovascular Nuclear Medicine. St. Louis: C. V. Mosby, 1979.

Sugimoto, T., *et al.* Quantitative effect of low coronary pressure on left ventricular performance. *Jap. Heart J.* 9:46, 1968.

Vedin, J. A., and Wilhelmsson, C. E. Beta receptor blocking agents in the secondary prevention of coronary heart disease. *Annu. Rev. Pharmacol. Toxicol.*, 23:29, 1983.

Wight, T. N. Proteoglycans in pathological conditions: Atherosclerosis. *Fed. Proc.*, 44:381, 1985.

Wunsford, M. D., *et al.* Calcium antagonists for acute ischemic heart disease. *Am. J. Cardiol.*, 55:116B, 1985.

Zahn, R., and Flaim, S. F. Calcium blocking drugs for angina pectoris. *Annu. Rev. Med.*, 33:465, 1982.

26

Cardiac Failure

Perhaps the most important ailment that must be treated by the physician is cardiac failure, which can result from any heart condition that reduces the ability of the heart to pump blood Usually the cause is decreased contractility of the myocardium resulting from diminished coronary blood flow, but failure to pump adequate quantities of blood can also be caused by damage to the heart valves, external pressure around the heart, vitamin deficiency, primary cardiac muscle disease, or any other abnormality that makes the heart a hypoeffective pump In the present chapter, we will discuss primarily cardiac failure caused by ischemic heart disease. In the following chapter we will discuss valvular and congenital heart disease

Definition of Cardiac Failure. The term cardiac failure means simply *failure of the heart, because of some malfunction of the heart itself, to pump enough blood* Cardiac failure may be manifest in either of two ways (1) by a decrease in cardiac output, or (2) by a damming of the blood in the veins behind the left or right heart—even though the cardiac output may be normal or at times even above normal, as we shall discuss later

Unilateral Versus Bilateral Cardiac Failure. Since the left and right sides of the heart are two separate pumping systems, it is possible for one of these to fail independently of the other For instance, unilateral failure can result from coronary thrombosis in one or the other of the ventricles However, cardiac failure from this cause occurs approximately 30 times as often in the left ventricle as in the right ventricle, so that there is a tendency among clinicians to view failure following myocardial infarction as almost always primarily left-sided. Occasionally, however, right-sided failure does occur with no left-sided failure at all, but this happens most frequently in persons with some congenital disease affecting primarily the right heart

In the first part of this chapter we will consider the whole heart failing as a unit and then return to the specific features of unilateral left- and right-sided failure.

DYNAMICS OF THE CIRCULATION IN CARDIAC FAILURE

ACUTE EFFECTS OF MODERATE CARDIAC FAILURE

If a heart suddenly becomes severely damaged in any way, such as by myocardial infarction, the pumping ability of the heart is immediately depressed. As a result, two essential effects occur: (a) reduced cardiac output and (b) damming of blood in the veins, resulting in increased systemic venous pressure. These two effects are shown graphically in Figure 26–1. This figure illustrates, first, a normal cardiac output curve, depicting the state of the circulation prior to the cardiac damage. Point A represents the normal state of the circulation, showing that the normal cardiac output under resting conditions is 5 liters per minuta and the right atrial pressure 0 mm Hg.

Immediately after the heart becomes damaged, the cardiac output curve becomes greatly reduced, falling to the lower, short-dashed curve at tha very bottom of the graph Within a few seconds after the acute heart attack, a new circulatory state is established at point B rather than point A, showing that the cardiac output has fallen to 2 liters per minute, about two-fifths normal, while the right atrial pressure has risen to 4 mm Hg because blood returning to the heart is dammed up in the right atrium This low cardiac output is still sufficient to sustain life, but it is likely to be associated with fainting. Fortunately, this acute stage lasts for only a few seconds because sympathetic reflexes occur immediately that can compensate to a great extent for tha damaged heart as follows.

Compensation for Acute Cardiac Failure by Sympathetic Reflexes. When the cardiac output falls precariously low, many of the different circulatory reflexes discussed in Chapter 21 are immediately activated The best known of these is the baroreceptor reflex, which is activated by diminished arterial pressure. It is probable that the chemoreceptor reflex, the central nervous system ischemic response, and possibly even reflexes originating in the damaged heart itself also contribute to a lesser extent to the nervous response. But whatever all the reflexes might be, the sympathetics become strongly stimulated within a few seconds, and the parasympathetics become reciprocally inhibited at the same time.

Strong sympathetic stimulation has two major effects on the circulation: first, on the heart itself, and, second, on the peripheral vasculature. Even the damaged myocardium responds to the sympathetic stimulation with increased force of contraction as well as greatly increased heart rate. If all the musculature is diffusely damaged, sympathetic stimulation strengthens this damaged musculature. Likewise, if part of the muscle is totally nonfunctional while part of it is still normal, the normal muscle is strongly stimulated by sympathetic stimulation, in this way compensating for the nonfunctional muscle. Thus, *the heart one way or another becomes a stronger pump, often as much as 100 per cent stronger, under the influence of the sympathetic impulses.* This effect is also illustrated in Figure 26–1, which

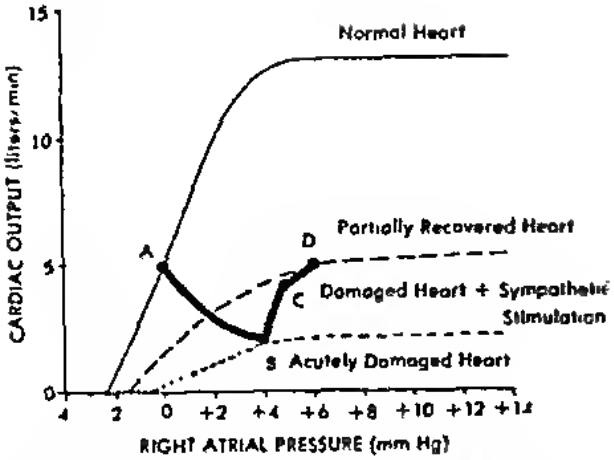


Figure 26–1. Progressive changes in the cardiac output curve following acute myocardial infarction. The cardiac output and right atrial pressure change progressively from point A to point D as explained in the text.

shows elevation of the cardiac output curve after sympathetic compensation (the dotted curve).

Sympathetic stimulation also increases the tendency for venous return, for it increases the tone of most of the blood vessels of the circulation, especially of the veins, *raising the mean systemic filling pressure* to 12 to 14 mm Hg, almost 100 per cent above normal. As will be recalled from the discussion in Chapter 23, this greatly increases the tendency for blood to flow back to the heart. Therefore, the damaged heart becomes primed with more inflowing blood than usual, and the right atrial pressure rises still further, which helps the heart to pump larger quantities of blood. Thus, in Figure 26–1 the new circulatory state is depicted by point C, showing a cardiac output of 4.2 liters per minute and a right atrial pressure of 5 mm Hg.

The sympathetic reflexes become maximally developed in about 30 seconds. Therefore, a person who has a sudden moderate heart attack might experience nothing more than cardiac pain and a few seconds of fainting. Shortly thereafter, with the aid of the sympathetic reflex compensations as described above, the cardiac output may return to a level entirely adequate to sustain the person who remains quiet, though the pain might persist.

THE CHRONIC STAGE OF FAILURE

After the first few minutes of an acute heart attack, a prolonged secondary state begins. This is characterized mainly by two events: (1) retention of fluid by the kidneys, and (2) progressive recovery of the heart itself over a period of several weeks to months, as was discussed in the previous chapter.

Renal Retention of Fluid

A low cardiac output has a profound effect on renal function, sometimes causing anuria when the cardiac output falls to as low as one-half to two-thirds normal. In general, the urinary output remains reduced as long as the cardiac output is significantly less than normal, and it usually does not return to normal after an acute heart attack until the cardiac output rises either all the way back to normal or almost to normal. This relationship of renal function to cardiac output is one of the most important of all the factors affecting the dynamics of the circulation in chronic cardiac failure.

Causes of Renal Retention of Fluid in Cardiac Failure. There are three known causes of the reduced renal output during cardiac failure, all of which are perhaps equally important but in different ways.

1. Decreased Glomerular Filtration. A decrease in cardiac output has a tendency to reduce the glomerular pressure in the kidneys because of (a) *reduced arterial pressure*, and (b) *intense sympathetic constriction of the afferent arterioles of the kidney*. As a consequence, except in the mildest degrees of heart failure, the glomerular filtration rate becomes less than normal. It will become evident from the discussion of kidney function in Chapter 35 that *even a very slight decrease in glomerular filtration often decreases urine output markedly*. When the cardiac output falls to about one-half normal, this factor alone can result in almost complete anuria.

2. Activation of the Renin-Angiotensin System and Increased Reabsorption of Water and Salt by the Kidneys. The reduced blood flow to the kidneys causes marked increase in renin output, and this in turn causes the formation of angiotensin by the mechanism described in Chapter 21. The angiotensin has a direct effect on the arterioles of the kidneys to decrease further the blood flow through the kidneys. Also, increased efferent arterial resistance decreases peritubular capillary pressure, thus promoting greatly increased reabsorption of both water and salt from the renal tubules. Therefore, the net loss of water and salt into the urine is greatly decreased, and the quantities of salt and water in the body fluids increase.

3. Increased Aldosterone Secretion. In the chronic stage of heart failure, large quantities of aldosterone are secreted by the adrenal cortex. This probably results mainly from the effect of angiotensin in stimulating aldosterone secretion. However, much of the increase in aldosterone secretion might also result from increased plasma potassium, because excess potassium is the most powerful stimulus known for aldosterone secretion, and the potassium concentration rises in response to reduced renal function in cardiac failure. Regardless of the cause of the increased aldosterone, it further increases the reabsorption of sodium from the renal tubules, and this in turn leads to a secondary increase in water reabsorption for two reasons: First, as the sodium is reabsorbed, it reduces the osmotic pressure in the tubules while increasing the osmotic pressure in the renal interstitial fluids; these changes promote osmosis of water into the blood. Second, the absorbed sodium increases the osmotic concentration of the extracellular fluid and elicits *antidiuretic hormone* secretion by the hypothalamic–posterior pituitary system, which is discussed in Chapter 36. The antidiuretic hormone then promotes increased tubular reabsorption of water.

The Beneficial Effects of Moderate Fluid Retention in Cardiac Failure. Though many cardiologists formerly considered fluid retention always to have a detrimental effect in cardiac failure, it is now known that a moderate increase in body fluid and blood volume is actually a very important factor helping to compensate for the diminished pumping ability of the heart. It does this by increasing the tendency for venous return. That is, the increased blood volume increases the venous return in two ways: First, it increases the mean systemic filling pressure, which increases *the pressure gradient*

for causing flow of blood toward the heart. Second, it distends the veins, which *reduces the venous resistance* and thereby allows increased ease of flow of blood to the heart.

If the heart is not too greatly damaged, this increased tendency for venous return can often fully compensate for the heart's diminished pumping ability—so much so, in fact, that if the heart's pumping ability is reduced to as little as 30 to 50 per cent of normal, still the increased venous return can often cause an entirely normal cardiac output.

Detrimental Effects of Excess Fluid Retention in the Severe Stages of Cardiac Failure. In contrast to the beneficial effects of moderate fluid retention in cardiac failure, in severe failure with extreme excesses of fluid retention the fluid then begins to have very serious physiological consequences, including overstretching of the heart thus weakening the heart still more, filtration of fluid into the lungs to cause pulmonary edema and consequent deoxygenation of the blood, and, often, development of extensive edema in all of the peripheral tissues of the body as well. These very detrimental effects of excessive fluid will be discussed in subsequent sections of the chapter

Recovery of the Myocardium Following Myocardial Infarction

After a heart becomes suddenly damaged as a result of myocardial infarction, the natural reparative processes of the body begin immediately to help restore normal cardiac function. Thus, a new collateral blood supply begins to penetrate the peripheral portions of the infarcted area, often completely restoring the muscle function. Also, the undamaged musculature hypertrophies, in this way offsetting much of the cardiac damage.

Obviously, the degree of recovery depends on the type of cardiac damage, and it varies from no recovery at all to almost complete recovery Ordinarily, after myocardial infarction the heart recovers rapidly during the first few days and weeks and will have achieved most of its final state of recovery within five to seven weeks, though some recovery continues for months

Cardiac Output Curve After Partial Recovery. The long-dashed curve of Figure 26–1 illustrates function of the partially recovered heart a week or so after the acute myocardial infarction By this time, considerable fluid also has been retained in the body, and the tendency for venous return has increased markedly; therefore, the right atrial pressure has risen even more. As a result, the state of the circulation is now changed from point C to point D, which represents a *normal* cardiac output of 5 liters per minute but with a right atrial pressure elevated to 6 mm Hg.

Since the cardiac output has returned to normal, renal output also will have returned to normal and no further fluid retention will occur. Therefore, except for the high right atrial pressure represented by point D in this figure, the person now has essentially normal cardiovascular dynamics *as long as he remains at rest.*

If the heart itself recovers to a significant extent and if adequate fluid retention occurs, the sympathetic stimulation gradually abates toward normal for the following reasons: The partial recovery of the heart can do the same thing for the cardiac curve as sympathetic stimulation, and fluid retention in the circulatory system can do the same thing for venous return as sympathetic stimulation. Thus, as these two factors develop, the fast pulse rate, cold skin, sweating, and pallor resulting from sympathetic stimulation in the acute stage of cardiac failure gradually disappear.

SUMMARY OF THE CHANGES THAT OCCUR FOLLOWING ACUTE CARDIAC FAILURE—"COMPENSATED HEART FAILURE"

To summarize the events discussed in the past few sections describing the dynamics of circulatory changes following an acute, moderate heart attack, we may divide the stages into (1) the instantaneous effect of the cardiac damage, (2) compensation by the sympathetic nervous system which occurs mainly within the first 30 seconds to 1 minute, and (3) chronic compensations resulting from partial cardiac recovery and renal retention of fluid. All these changes are shown graphically by the very heavy line in Figure 26–1. The progression of this line shows: the normal state of the circulation (point A), the state a few seconds after the heart attack but before sympathetic reflexes have occurred (point B), the rise in cardiac output toward normal caused by sympathetic stimulation (point C), and final return of the cardiac output to normal following several days to several weeks of cardiac recovery and fluid retention (point D). This final state is called *compensated heart failure.*

Compensated Heart Failure. Note especially in Figure 26–1 that the pumping ability of the heart, as depicted by the cardiac output curve, is still depressed to less than one-half normal. This illustrates that factors that increase the right atrial pressure (principally retention of fluid) can maintain the cardiac output at a normal level despite continued weakness of the heart itself. However, one of the results of chronic cardiac weakness is this chronic increase in right atrial pressure itself; in Figure 26–1 it is shown to be 6 mm Hg. Many persons, especially in old age, have completely normal resting cardiac outputs but mildly to moderately elevated right atrial pressures because of compensated heart failure. These persons may not know that they have cardiac damage because the damage more often than not has occurred a little at a time, and the compensation has occurred concurrently with the progressive stages of damage.

Reduced Cardiac Reserve in Compensated Failure. The slight elevation of right atrial pressure has little harmful effect on the circulatory system (perhaps a slightly enlarged liver and tendency toward ankle edema). However, the person with compensated heart failure certainly does not have a normal circulatory system, for should that person try to exercise strongly or should any other stress be placed on the circulatory system, such as might occur in some disease condition, the heart would be unable to respond normally. Instead, the person becomes very weak and often develops severe dyspnea because the circulation cannot deliver the needed oxygen to the nervous system and also cannot remove the carbon dioxide In addition, the sympathetic system becomes excessively stimulated in a futile attempt to excite the circulation to a greater level of cardiac output—the heart rate becomes very rapid, the skin pallid, and sweat beads often appear.

To state this another way, the normal person has far greater *cardiac reserve* than the person with compensated heart disease. Cardiac reserve is discussed at greater length later in the chapter.

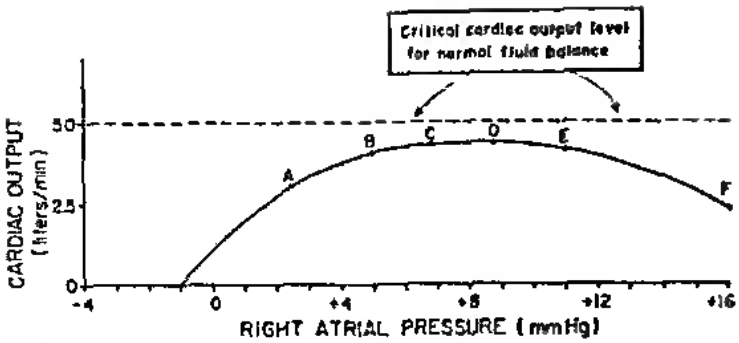


Figure 26–2. Greatly depressed cardiac output curve that indicates decompensated heart disease. Progressive fluid retention raises the right atrial pressure and the cardiac output progresses from point A to point F, as explained in the text.

DYNAMICS OF SEVERE CARDIAC FAILURE—DECOMPENSATED HEART FAILURE

If the heart becomes severely damaged, then no amount of compensation, either by sympathetic nervous reflexes or by fluid retention, can make this weakened heart pump a normal cardiac output. As a consequence, the cardiac output cannot rise to a high enough value to bring about return of normal renal function. Fluid continues to be retained, the person develops progressively more and more edema, and this state of events eventually leads to death. This is called *decompensated heart failure*. The main basis of decompensated heart failure is *failure of the heart to pump sufficient blood to make the kidneys function adequately*.

Analysis of Decompensated Heart Failure. Figure 26–2 illustrates a greatly depressed cardiac output curve, depicting the function of a heart that has become extremely weakened and cannot be strengthened. Point A on this curve represents the approximate state of the circulation before any compensation has occurred, and point B the state after the first few minutes of acute compensation by sympathetic stimulation, as described above. At point B the cardiac output has risen to 4 liters per minute and the right atrial pressure to 5 mm Hg. The person appears to be in reasonably good condition, but this state will not remain static for the following reason. The cardiac output has not risen quite high enough to cause adequate kidney excretion of fluid, therefore, fluid retention continues unabated and can eventually be the cause of death. These events can be explained quantitatively in the following way.

Note the dashed line at a cardiac output level of 5 liters in Figure 26–2. This is the critical cardiac output level that is required to allow the kidneys to re-establish normal fluid balance. At any cardiac output below this level, all the fluid-retaining mechanisms discussed in the earlier section remain in play, and the body fluid volumes will increase progressively. Because of this progressive increase in fluid volume, the mean systemic filling pressure continues to rise, and this forces progressively increasing quantities of blood into the right atrium, thus increasing the right atrial pressure. After a day or so, the state of the circulation changes from point B in Figure 26–2 to point C—the right atrial pressure rising to 7 mm Hg and the cardiac output to 4.2 liters per minute. However, note again that the cardiac output is still not high enough to cause normal renal output of fluid; therefore, fluid continues to be retained, and after another day or so the right atrial re rises to 9 mm Hg, and the circulatory state es that depicted by point D. Still, the cardiac is not enough to establish normal fluid balance.

After another few days the right atrial pressure has risen still further, but by now the cardiac function curve is beginning to decline toward a lower level. This decline is caused by overstretch of the heart, edema of the heart muscle, and other factors that diminish the pumping performance of the heart. It is especially evident, however, that further retention of fluid from then on will be more detrimental than beneficial to the circulation. Still, the cardiac output is not high enough to bring about normal renal function, and fluid retention not only continues but actually accelerates because of the falling cardiac output. Consequently, within a few days the state of the circulation has reached point F on the curve, with the cardiac output less than 2.5 liters per minute and the right atrial pressure 16 mm Hg. This is a state that now has approached or reached incompatibility with life, and the patient dies. This state of heart failure in which the failure continues to worsen is called *decompensated heart failure*.

Thus, one can see from this analysis that failure of the cardiac output ever to rise to the critical level required for normal renal function results in (a) progressive retention of fluid, which causes (b) progressive elevation of the mean systemic filling pressure, and (c) progressive elevation of the right atrial pressure until finally the heart is so overstretched or so edematous that it becomes unable to pump even moderate quantities of blood, and, therefore, fails completely. Clinically, one detects this serious condition of decompensation principally by the progressive edema, especially edema of the lungs, which leads to bubbling rales in the lungs and dyspnea (air hunger). All clinicians know that failure to institute appropriate therapy when this state of events occurs leads to rapid death.

Myocardial Deterioration in Decompensated Failure. In addition to the severe problem of progressive fluid retention in decompensated heart failure, another important complicating problem is progressive deterioration of the myocardium itself. The basic causes of this are probably such factors as (1) diminished blood supply to the cardiac muscle, (2) overstretching of the muscle, (3) edema of the muscle, (4) growth of fibrous tissue in the myocardium, (5) overworking of the remaining functional musculature, and so on. But, regardless of the cause or causes of the deterioration, some of the important results are the following.

First, the contractility of the heart muscle itself diminishes.

Second, the calcium mechanism for causing excitation-contraction coupling is depressed. This results partly from the fact that the longitudinal tubules of the sarcoplasmic reticulum fail to accumulate normal amounts of calcium. Therefore, when the muscle is stimulated,

not enough calcium is available to cause full contraction In addition, the troponin-tropomyosin system for initiating interaction between the actin and myosin filaments seems to respond less intensely to the calcium that is available. These effects also explain why digitalis is often very effective in improving the contractile strength of the failing heart, because digitalis has a specific function to increase the buildup of calcium ions in cardiac muscle.

Third, the quantity of stored norepinephrine in the sympathetic nerve endings decreases markedly in the failing myocardium And the rate of secretion of norepinephrine in response to sympathetic stimuli is correspondingly reduced. Therefore, the failing myocardium responds more and more poorly to sympathetic stimulation This means that the heart loses one of its very important stimulating mechanisms

One can well understand that such cardiac deterioration occurring in the late stages of cardiac failure can lead to still more failure Once this begins, it has a tendency to perpetuate itself and even to become a vicious circle. Therefore, intensive therapy in this stage of heart disease is essential if the person is to survive

Treatment of Decompensation. The two ways in which the decompensation process can often be stopped are (1) by strengthening the heart in any one of several ways, especially by administration of a cardiotonic drug, such as digitalis, so that it can pump adequate quantities of blood to make the kidneys function normally again, or (2) by administering diuretics and reducing water and salt intake, which brings about a balance between fluid intake and output despite the low cardiac output

Both methods stop the decompensation process by re-establishing normal fluid balance so that at least as much fluid leaves the body as enters it

Mechanism of Action of the Cardiotonic Drugs. Cardiotonic drugs, such as digitalis, have little effect on increasing the contractile strength of normal cardiac muscle. On the other hand, when administered to a person with chronic cardiac failure these same drugs can sometimes increase the strength of the failing myocardium as much as 50 to 100 per cent Therefore, they are the mainstay of therapy in chronic heart failure

The way in which digitalis and the other cardiotonic glycosides strengthen heart contraction is believed to be by increasing the quantity of calcium ions in the muscle fibers. It was pointed out earlier that the cardiac muscle sarcoplasmic reticulum fails to accumulate normal quantities of calcium in severe heart failure and therefore cannot release enough calcium to cause full contraction of the muscle Consequently, this accumulation of extra calcium in the muscle fibers at least partially corrects this problem.

UNILATERAL CARDIAC FAILURE

In the discussions thus far in this chapter, we have considered failure of the heart as a whole. Yet in a large number of patients, especially those with early acute failure, left-sided failure predominates over right-sided failure, and in rare instances, especially in congenital heart disease, the right side may fail without significant failure of the left side. Therefore, we now need to discuss the special features of unilateral failure.

UNILATERAL LEFT HEART FAILURE

When the left side of the heart fails without concomitant failure of the right side, blood continues to be pumped into the lungs with usual right heart vigor while it is not pumped adequately out of the lungs into the systemic circulation As a result, the *mean pulmonary filling pressure* rises because of shift of large volumes of blood from the systemic circulation into the *pulmonary circulation.*

As the volume of blood in the lungs increases, the pulmonary vessels enlarge, and, if the pulmonary capillary pressure rises above 28 mm Hg, that is, above the colloid osmotic pressure of the plasma, fluid begins to filter out of the capillaries into the interstitial spaces and alveoli, resulting in pulmonary edema

Thus, among the most important problems of left heart failure are *pulmonary vascular congestion* and *pulmonary edema*, which are discussed in detail in Chapters 24 and 31 in relation to the pulmonary circulation and capillary dynamics As long as the pulmonary capillary pressure remains less than the normal colloid osmotic pressure of the blood, ranging between 23 and 33 mm Hg, the lungs remain "dry." But even a few millimeters' rise in capillary pressure above this critical level, when this occurs acutely, causes progressive transudation of fluid into the interstitial spaces and alveoli, leading rapidly to death. Pulmonary edema can occur so rapidly that it can cause death after only 20 to 30 minutes of severe acute left heart failure. (However, in chronic left heart failure, the lymphatic vessels of the lungs enlarge, so that lethal pulmonary edema sometimes then will not occur even with left atrial pressures as high as 40 mm Hg.)

Course of Events for Several Days After Acute Left Heart Failure

During the several days after the onset of left heart failure, one additional feature must be added to the acute picture This is retention of fluid resulting from reduced renal function. In moderate left heart failure, the pulmonary capillary pressure sometimes rises only to 15 to 20 mm Hg during the first days of the failure, as illustrated in Figure 26–3, not enough to cause pulmonary edema. Yet following retention of fluid for the next few days, the blood volume increases, and more blood is pumped into the lungs by the right ventricle. Then, the pulmonary capillary pressure rises still more, often rising above the plasma colloid osmotic pressure, resulting in severe pulmonary edema as shown in the figure Indeed, this is a common occurrence: the patient suddenly develops severe pulmonary edema a few days to a week or so after the acute attack and dies a respiratory death, not a death resulting from diminished cardiac output.

UNILATERAL RIGHT HEART FAILURE

In unilateral right heart failure, blood is not pumped adequately from the systemic circulation into the lungs. Therefore, blood shifts from the lungs into the systemic circulation. However, the effect of this on the systemic circulation is hardly noticeable for the following reason: The total amount of blood in the lungs is only about one ninth that in the systemic circulation. Therefore, even

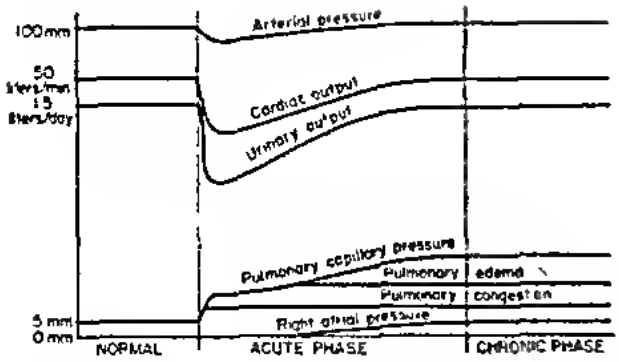


Figure 26–3. The overall effects, acute and chronic of left-sided heart failure

in severe *acute* right heart failure, the systemic blood volume increases only a few per cent because of blood shift from the lungs, and this is not sufficient to cause significant systemic congestion

Low Cardiac Output in Acute Right Heart Failure. On the other hand, acute right heart failure can cause greater depression of the cardiac output than acute left heart failure of the same degree This again stems from the far greater compliance of the systemic circulation than of the pulmonary circulation Not enough blood can transfer from the lungs into the systemic vessels to raise the systemic pressures to a very high level, and these pressures are not enough to make the weakened right ventricle pump adequate quantities of blood Therefore, in those rare conditions in which the right heart does fail acutely, the cardiac output falls greatly, often leading to such low cardiac output that death ensues rapidly

The Chronic Stage of Unilateral Right Heart Failure. The chronic stage of unilateral right heart failure is much the same as that discussed earlier for the entire heart The depressed cardiac output results in progressively more and more retention of fluid by the kidneys until the cardiac output either rises back nearly to normal or until the person goes into decompensation and dies Figure 26–4 illustrates the progressive changes that occur in chronic right-sided heart failure, showing a gradual return of cardiac output to normal or near normal, a return of arterial pressure and urinary output to normal, but also progressive development of peripheral congestion and edema—yet no pulmonary edema

"HIGH CARDIAC OUTPUT FAILURE"—OVERLOADING OF THE HEART

A condition called "high cardiac output failure" frequently occurs in persons who have cardiac outputs much higher than normal but who also have signs of cardiac failure because the left or right or both atrial pressures are much too high Sometimes there is also accumulation of edema The true problem is often not failure of the pumping ability of the heart but instead *overloading of the heart with too much venous return.* Most often this is caused by some circulatory abnormality that decreases the systemic vascular resistance, such as

(1) *Arteriovenous fistula,* in which blood is shunted directly from the arteries to the veins. This greatly increases the venous return and frequently causes congestive symptoms similar to those of typical heart failure, even though the heart is normal and is actually pumping a tremendously increased cardiac output

(2) *Beriberi heart disease,* which is caused by lack of the B vitamins, especially of thiamine This results in profound systemic vasodilatation, sometimes decreasing the systemic resistance to as little as one-half normal The heart is usually weakened as well, yet, even with the weakness of the heart, the cardiac output is still as much as two times normal Therefore, severe congestive symptoms occur both because of the excess venous return and because of the weakened heart, so that there is *both excess loading of the heart and true heart failure.*

(3) *Thyrotoxicosis,* which, because of its stimulatory effect on body metabolism, causes generalized systemic vasodilatation Here again, the thyrotoxicosis frequently weakens the heart as well, so that there is often both overloading of the heart and true heart failure, both occurring simultaneously

LOW CARDIAC OUTPUT IN HEART FAILURE—"CARDIOGENIC SHOCK"

In the description of acute heart failure, it was pointed out that the cardiac output can fall very low immediately after heart damage occurs This obviously leads to greatly diminished blood flow throughout the body and can lead to a typical picture of circulatory shock as described in Chapter 28 This means, simply, that the cardiac output falls too low to supply the body with adequate blood flow As a result, the tissues deteriorate rapidly and death ensues Sometimes death comes in less than an hour; at other times, it comes over a period of several days The circulatory shock that is caused by inadequate cardiac pumping is called *cardiogenic shock* or *cardiac shock,* and it is sometimes also called the *pump failure syndrome*

Cardiogenic shock is extremely important to the clinician because approximately one tenth of all patients who have acute myocardial infarction die because of the low cardiac output before the physiologic compensatory measures can come into play to save the person's life Once cardiac shock has become well-established after myocardial infarction, all the typical events occur that also occur in the late stages of other types of circulatory shock, as described in Chapter 28, especially rapid deterioration of almost all bodily functions

Vicious Circle of Cardiac Deterioration in Car-

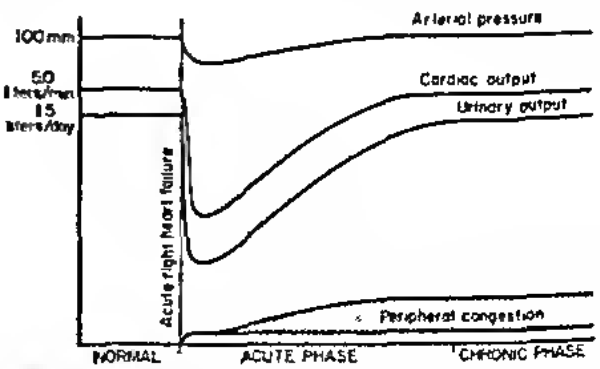


Figure 26–4. The overall effects, acute and chronic, of right sided heart failure

diogenic Shock. The discussion of circulatory shock in Chapter 28 will emphasize the tendency for the heart itself to become progressively more damaged when its coronary blood supply is reduced during the course of shock. That is, the low arterial pressure that occurs during shock reduces the coronary supply, which makes the heart still weaker, which makes the shock still worse, the process eventually becoming a vicious circle of cardiac deterioration. In cardiogenic shock caused by myocardial infarction, this problem is greatly compounded by the already existing coronary thrombosis For instance, in a normal heart, the arterial pressure usually must be reduced below about 45 mm Hg before cardiac deterioration sets in. However, in a heart that already has a blocked major coronary vessel, deterioration will set in when the arterial pressure falls as low as 80 to 90 mm Hg. In other words, even the minutest amount of fall in arterial pressure can set off a vicious circle of cardiac deterioration following myocardial infarction, though this circle will occur in a normal heart only in the very late stages of shock For this reason, in treating myocardial infarction it is extremely important to prevent even short periods of hypotension

Physiology of Treatment. Often a patient dies of cardiogenic shock before the various compensatory processes can return the cardiac output to a life-sustaining level. Therefore, treatment of this condition is one of the most important problems in the management of acute heart attacks Immediate digitalization of the heart is often employed for strengthening the heart if the remaining functioning ventricular muscle shows signs of deterioration Also, infusion of whole blood, plasma, or a blood pressure-raising drug is used to sustain the arterial pressure If the arterial pressure can be raised high enough, the coronary blood flow can often be elevated to a high enough value to prevent the vicious circle of deterioration until appropriate compensatory mechanisms in the body can correct the shock. Even with the best therapy, however, once the shock syndrome has begun, with the arterial pressure remaining as much as 20 mm Hg below normal for as long as an hour, 85 per cent of the patients die

Some success has been achieved recently in saving the lives of patients with cardiogenic shock using one of the following two procedures (1) surgical removal of the clot in the coronary artery, often combined with coronary bypass graft, or (2) catheterizing the blocked coronary artery and infusing streptokinase, an enzyme that causes dissolution of the clot The results are occasionally astounding when the procedure is instituted within the first hour but of little if any benefit after three hours.

EDEMA IN PATIENTS WITH CARDIAC FAILURE

Much of the preceding discussion has emphasized the edema that develops in cardiac failure. On first thought one would think that it should be easy to understand the cause of edema when the heart fails. One's natural logic directs that when the heart fails, blood becomes dammed up behind the heart, thereby increasing the venous and capillary pressures until fluid transudes out of the capillaries into the tissues. However, peripheral edema will not occur until large quantities of water and salt are retained by the kidneys. Consequently, there are special quirks of logic that must be explained before one can understand some of the problems related to edema in cardiac failure.

Inability of Acute Cardiac Failure to Cause Peripheral Edema. Though acute left heart failure can cause terrific congestion of the lungs, with rapid development of pulmonary edema, acute heart failure of any type never causes immediate development of peripheral edema. This can be explained best by referring to Figure 26–5. When a previously normal heart acutely fails as a pump, the arterial pressure falls and the right atrial pressure rises. The two pressures approach each other at an equilibrium value of about 13 mm Hg. It is obvious that capillary pressure must also fall from its normal value of 17 mm Hg to the equilibrium pressure of 13 mm Hg Thus, *severe acute cardiac failure actually causes a fall in peripheral capillary pressure rather than a rise.* And animal experiments as well as experience in human beings show that acute cardiac failure does not cause immediate development of peripheral edema In fact, during the first hours fluid is actually absorbed into the capillaries from the interstitial spaces because the colloid osmotic pressure of the plasma becomes more important in absorbing fluid when the capillary pressure falls

Role of the Kidneys in Cardiac Edema. After the first day or so, peripheral edema does begin to occur, but principally *because of fluid retention by the kidneys.* The retention of fluid increases the mean systemic filling pressure, resulting in increased tendency for blood to return to the heart This now elevates the right atrial pressure to a still higher value, and it also returns the arterial pressure to normal Therefore, the rising right atrial pressure *now causes the capillary pressure to rise markedly,* thus causing loss of fluid into the tissues and development of severe edema

The mechanisms of fluid retention by the kidneys were discussed earlier in the chapter, but these are so important that they need to be related here for our present discussion (1) decreased glomerular pressure because of slightly decreased arterial pressure and especially because of sympathetic constriction of the afferent arterioles, (2) formation of angiotensin, which causes efferent arteriolar constriction with resultant excessive reabsorption of water and salt by the renal tubules; (3)

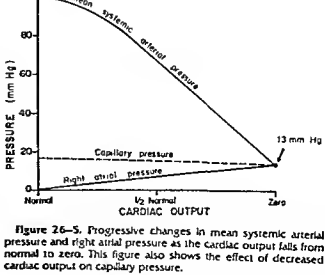


Figure 26–5. Progressive changes in mean systemic arterial pressure and right atrial pressure as the cardiac output falls from normal to zero. This figure also shows the effect of decreased cardiac output on capillary pressure.

marked increase in aldosterone secretion, which also causes greatly increased salt and water reabsorption by the renal tubules, and (4) increased ADH secretion, which markedly increases water reabsorption by the renal tubules

Acute Pulmonary Edema In Chronic Heart Failure—A Lethal Vicious Circle

A frequent cause of death is acute pulmonary edema occurring in patients who have had chronic heart failure for a long time Sometimes this edema is caused by a new myocardial infarction, but in many instances no new pathology is found in the heart

When acute pulmonary edema occurs in a person without new cardiac damage, it usually is set off by some temporary overload of the heart, such as might result from a bout of heavy exercise, some emotional experiences, or even a severe cold This acute pulmonary edema is believed to result from the following vicious circle

1 A temporarily increased load on the already weak left ventricle results from increased venous return from the peripheral circulation Because of the limited pumping capacity of the left heart, blood begins to dam up in the lungs

2. The increased blood in the lungs increases the pulmonary capillary pressure, and a small amount of fluid begins to transude into the lung tissues and alveoli

3 The increased fluid in the lungs diminishes the degree of oxygenation of the blood

4 The decreased oxygen in the blood further weakens the heart and also causes peripheral vasodilation

5 The peripheral vasodilatation increases venous return from the peripheral circulation still more

6 The increased venous return further increases the damming of the blood in the lungs, leading to still more transudation of fluid, more arterial oxygen desaturation, more venous return, and so forth Thus, a vicious circle has been established

Once this vicious circle has proceeded beyond a certain critical point, it will continue until death of the patient unless heroic therapeutic measures are employed The types of heroic therapeutic measures that can reverse the process and save the life of the patient include

a *Putting tourniquets on all four limbs to sequester* much of the blood in the veins and therefore to decrease the work load on the left heart

b Bleeding the patient

c Giving a rapidly acting diuretic such as furosemide to cause rapid loss of fluid from the body

d Giving the patient pure oxygen to breathe to reverse the blood desaturation, the heart deterioration, and peripheral vasodilatation

e Giving the patient a rapidly acting cardiotonic drug such as ouabain to strengthen the heart

Unfortunately, this vicious circle of acute pulmonary edema can proceed so rapidly that death can occur within 30 minutes to an hour Therefore, any procedure that is to be successful must be instituted immediately

PHYSIOLOGICAL CLASSIFICATION OF CARDIAC FAILURE

From these discussions, it is apparent that the symptoms of cardiac failure fall into the following three physiological classifications:

Low cardiac output
Pulmonary congestion
Systemic congestion.

Low cardiac output usually occurs immediately af a heart attack. If the attack is mainly right-sided, t may be the only symptom. If the acute heart attack mainly left-sided, concurrent pulmonary congestion most always occurs along with the low cardiac outp but the pulmonary congestion may be mild (with pulmonary edema) until after considerable fluid l been retained by the kidneys. Thus, low cardiac out may be the only significant clinical effect observed many persons who have sudden heart attacks T results in the following symptoms

1 Generalized weakness

2 Fainting

3 Symptoms of increased sympathetic activity s as high heart rate, thready pulse, cold skin, sweati and so forth

Pulmonary congestion may be the only effect in tients with pure left-sided, *chronic* heart failure, beca in the chronic stage enough fluid will have been retai to return the cardiac output to normal despite the we left ventricle—but this occurs at the expense of grea elevated pulmonary vascular pressures And, since t right heart is not failing, pulmonary congestive syn toms alone can occur with essentially no syster congestion nor low cardiac output

Systemic congestion alone can occur in pure rig sided *chronic* heart failure In this condition there is pulmonary congestion, and if sufficient fluids have be retained in the blood to prime the heart sufficiently t heart may pump a normal cardiac output

Obviously, all these classes of heart failure can oc together or in any combination

CARDIAC RESERVE

The maximum percentage that the cardiac output c increase above normal is called the *cardiac reser* Thus, in the normal young adult the cardiac reserve 300 to 400 per cent In the athletically trained person is occasionally as high as 500 to 600 per cent, where in the asthenic person it may be as low as 200 per ce As an example, during severe exercise the cardiac outp of the normal healthy young adult can rise to about fi times normal, this is an increase above normal of 4 per cent—that is, a cardiac reserve of 400 per cent

Any factor that prevents the heart from pumpi blood satisfactorily decreases the cardiac reserve Th can result from ischemic heart disease, primary my cardial disease, vitamin deficiency, damage to the my cardium, valvular heart disease, and many other facto some of which are illustrated in Figure 26–6

Diagnosis of Low Cardiac Reserve—The Exerci Test. So long as people with low cardiac reserve rema in a state of rest, they probably will not know that th have heart disease. However, a diagnosis of low cardi reserve can usually be made easily by requiring a pers to exercise either on a treadmill or by walking up a down steps. The increased load on the heart rapidly us up the small amount of reserve that is available, a the cardiac output fails to rise high enough to susta the body's new level of activity The acute effects are

1 Immediate and sometimes extreme shortness breath (dyspnea) resulting from the heart's not pumpi

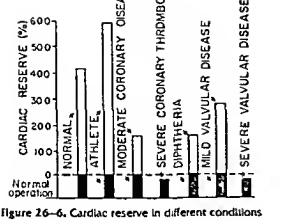


Figure 26–6. Cardiac reserve in different conditions

sufficient blood to the tissues, thereby causing tissue ischemia and creating a sensation of air hunger

2. Extreme muscle fatigue resulting from muscle ischemia, thus limiting the person's ability to continue with the exercise.

3. Excessive increase in heart rate because the nervous reflexes overreact in an attempt to overcome the inadequate cardiac output

Therefore, exercise tests are part of the armamentarium of the cardiologist These tests take the place of cardiac output measurements that unfortunately cannot be made with ease in most clinical settings

APPENDIX
A QUANTITATIVE GRAPHICAL METHOD FOR ANALYSIS OF CARDIAC FAILURE

Though it is possible to understand most of the general principles of cardiac failure using mainly qualitative logic, as we have done thus far in this chapter, one can grasp the importance of the different factors in cardiac failure with far greater depth by using quantitative approaches. One such approach is the graphical method for analysis of cardiac output regulation that was introduced in Chapter 23 In the following few paragraphs, we will analyze several aspects of cardiac failure utilizing this graphical technique.

Graphical Analysis of Acute Heart Failure and of Chronic Compensation

Figure 26–7 illustrates cardiac output and venous return curves for different states of the heart and of the peripheral circulation. The two *solid curves* are (1) the normal cardiac output curve, and (2) the normal venous return curve. As pointed out in Chapter 23, there is only one point on each of these two curves at which the circulatory system can operate so long as these two curves remain normal. This point is where the two curves cross at point A. Therefore, the normal state of the circulation is a cardiac output and venous return of 5 liters per minute and a right atrial pressure of 0 mm Hg.

Effect of Acute Heart Attack. During the first few seconds after a moderately severe heart attack, the cardiac output curve falls to the lowermost *long-dashed curve* During this few seconds, the venous return curve still has not changed because the peripheral circulatory system is still operating normally. Therefore, the new state of the circulation is depicted by *point B*, where the new cardiac output curve crosses the normal venous return curve Thus the right atrial pressure rises to 4 mm Hg, whereas the cardiac output falls to 2 liters per minute

Effect of Sympathetic Reflexes. Within the next 30 seconds, the sympathetic reflexes become very active These affect both curves, raising them to the *short-dashed curves.* Sympathetic stimulation can increase the plateau level of the cardiac output curve as much as 100 per cent. It can also increase the mean systemic filling pressure (depicted by the point where the venous return curve crosses the zero venous return axis) by several millimeters Hg—in this figure from a normal value of 7 mm Hg up to 10 5 mm Hg This increase in mean systemic filling pressure shifts the entire venous return curve to the right and upward. The new cardiac output and venous return curves now equilibrate at point C, that is, at a right atrial pressure of 5 mm and at a cardiac output of 4 liters per minute.

Compensation During the Next Few Days. During the ensuing week, the cardiac output and venous return curves rise to the *dot-dashed curves* because of (1) some recovery of the heart and (2) renal retention of salt and water, which raises the mean systemic filling pressure still further—this time up to 12 mm Hg The two new curves, the dot-dashed curves, now equilibrate at *point D* Thus, the cardiac output has now returned entirely to normal The right atrial pressure, however, has risen still further to 6 mm Hg. Because the cardiac output is now normal, renal output is also normal, so that a new state of equilibrated fluid balance has been achieved. Therefore, the circulatory system will continue to function at point D and will remain stable, with a normal cardiac output and an elevated right atrial pressure, until some additional extrinsic factor changes either the cardiac output curve or the venous return curve.

Utilizing this technique for analysis, one can see especially the importance of fluid retention and how it eventually leads to a new stable state of the circulation; one can also see the interrelationship between the mean systemic filling pressure and cardiac pumping at various degrees of cardiac failure.

Note especially that the events described in Figure 26–7 are the same as those presented in Figure 26–1 but presented in a more quantitative manner.

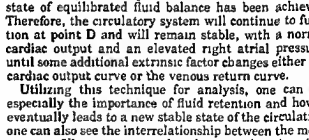


Figure 26–7. The progressive changes in cardiac output and right atrial pressure during different stages of cardiac failure.

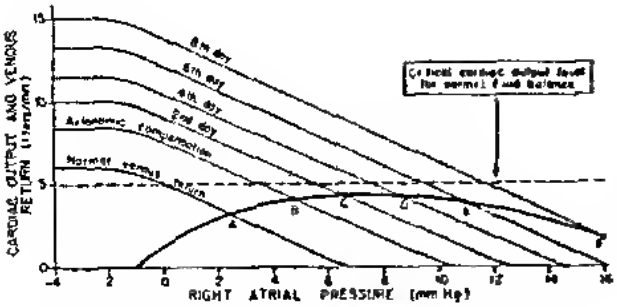


Figure 26–8. Graphical analysis of decompensated heart disease, showing progressive shift of the venous return curve to the right as a result of fluid retention.

Graphical Analysis of "Decompensated" Cardiac Failure

The cardiac output curve in Figure 26–8 is the same as that in Figure 26–2, a very low curve that in this case has already reached a degree of recovery as great as this heart can achieve. In this figure we have added venous return curves that occur during successive days following the acute fall of the cardiac output curve to this very low level. At *point A* the curve equates with the normal venous return curve to give a cardiac output of approximately 3 liters per minute. However, stimulation of the sympathetic nervous system, caused by this low cardiac output, increases the mean systemic filling pressure within 30 seconds from 7 to 10.5 mm Hg and shifts the venous return curve upward and to the right to give the curve labeled "autonomic compensation." Thus, the new venous return curve equates with the cardiac output curve at *point B*. The cardiac output has been improved up to a level of 4 liters per minute but at the expense of an additional rise in right atrial pressure to 5 mm Hg.

The cardiac output of 4 liters per minute is still too low to cause the kidneys to function normally. Therefore, fluid continues to be retained, the mean systemic filling pressure rises from 10.5 almost to 13 mm Hg, and the venous return curve becomes that labeled "second day." This equilibrates with the cardiac output curve at *point C*, and the cardiac output rises to 4.2 liters per minute while the right atrial pressure rises to 7 mm Hg.

During the succeeding days the cardiac output never rises high enough to re-establish normal fluid balance. Therefore, fluid continues to be retained, the mean systemic filling pressure continues to rise, the venous return curve continues to shift to the right, and the equilibrium point between the venous return curve and the cardiac output curve shifts progressively to *point D*, to *point E*, and finally to *point F*. The equilibration process is now on the downslope of the cardiac output curve so that further retention of fluid causes only a detrimental effect on cardiac output. Therefore, the condition accelerates in a downhill direction until death occurs.

Thus, the process of "decompensation" results from the fact that the cardiac output curve never rises to the critical level of 5 liters per minute required to re-establish normal balance between fluid input and output.

Treatment of Decompensated Heart Disease with Digitalis. Let us assume that the stage of decompensation has already reached point E in Figure 26–8, and let us proceed to the same point E in Figure 26–9. At this point, digitalis is given to strengthen the heart and to raise the cardiac output curve to the level of the heavy curve in Figure 26–9. This does not immediately change the venous return curve. Therefore, the new cardiac output curve equates with the venous return curve at *point G*. The cardiac output is now 5.7 liters per minute, a value greater than the critical level of 5 liters required for normal fluid balance. Therefore, the kidneys eliminate far more fluid than normally, causing diuresis, a well known effect of digitalis.

The progressive loss of fluid over a period of several days reduces the mean systemic filling pressure back down to 11.5 mm Hg, and the new venous return curve becomes the heavy curve labeled "several days later." This curve equates with the cardiac output curve of the digitalized heart at *point H*, at an output of 5 liters per minute and a right atrial pressure of 4.6 mm Hg. This cardiac output is precisely that required for normal fluid balance. Therefore, no additional fluid will be lost, and none will be gained. Consequently, the circulatory system has now stablized, or, in other words, the decompensation of the heart failure has been "compensated." To state this another way, the final steady-state condition of the circulation is defined by the crossing point of three different curves: the cardiac output curve, the venous return curve, and the critical level for normal fluid balance. The compensatory mechanisms automatically stablize the circulation when all three curves cross at the same point.

Graphical Analysis of High Cardiac Output Failure. Figure 26–10 gives an analysis of two types of high cardiac output failure. One of these is caused by an arteriovenous fistula that overloads the heart because of excessive venous return, even though the pumping capability of the heart itself is not depressed. The other is caused by the disease beriberi, in which the venous return is greatly increased because of diminished systemic vascular resistance, but at the same time the pumping capability of the heart is also depressed.

Arteriovenous Fistula. The heavy black curves of the figure depict the normal cardiac output and normal venous return curve. These equate with each other at *point A*, which depicts a *normal* cardiac output of 5 liters per minute and a normal right atrial pressure of 0 mm Hg.

Now let us assume that systemic resistance (the total peripheral resistance) becomes greatly decreased be-

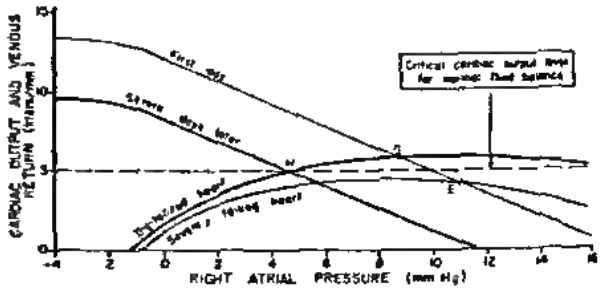


Figure 26–9. Treatment of decompensated heart disease, showing the effect of digitalis in elevating the cardiac output curve, this in turn causing progressive shift of the venous return curve to the left.

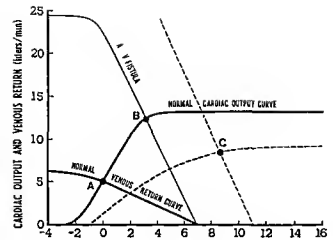


Figure 26–10. Graphical analysis of the two types of factors that can cause high cardiac output failure

cause of opening a very large arteriovenous fistula. The venous return curve rotates upward to give the curve labeled "A-V fistula." This venous return curve equates with the cardiac output curve at *point B*, with a cardiac output of 12.5 liters per minute and a right atrial pressure of 3 mm Hg. Thus, the cardiac output has become greatly elevated, the right atrial pressure is slightly elevated, and there are mild signs of peripheral congestion. If the person attempts to exercise, he or she will have very little cardiac reserve because the heart is already being used almost to maximum capacity simply to pump the extra blood through the arteriovenous fistula. Therefore, this condition resembles a failure condition and is called "high output failure," but in reality the heart is simply overloaded by excess venous return.

Beriberi. The dashed lines in Figure 26–10 illustrate the changes in the cardiac output and venous return curves caused by beriberi. The decreased level of the cardiac output curve is caused by weakening of the heart because of the avitaminosis (mainly lack of thiamine) that causes the beriberi syndrome. The weakening of the heart has decreased the blood flow to the kidneys. Therefore, the kidneys have retained a large amount of fluid, which in turn has increased the mean systemic filling pressure (represented by the point where the venous return curve intersects with the zero cardiac output level) from the normal value of 7 mm Hg up to 11 mm Hg. This has shifted the venous return curve to the right. Finally, the venous return curve has rotated upward because the avitaminosis has dilated the peripheral blood vessels, as was explained in Chapter 20.

Therefore, the two dashed curves intersect with each other at *point C*, which describes the circulatory condition in beriberi, with a right atrial pressure in this instance of 9 mm Hg and a cardiac output about 65 per cent above normal—this very high cardiac output occurs despite the weak heart, as illustrated by the depressed plateau level of the cardiac output curve.

REFERENCES

Barger, A C · The kidney in congestive heart failure *Circulation, 21* 124, 1960

Bradley, R D · Studies in Acute Heart Failure. London, Edward Arnold, 1977

Braunwald, E · Heart failure—pathophysiological considerations *In* Dickinson, C J, and Marks, J (eds.) Developments in Cardiovascular Medicine Lancaster, England, MTP Press, 1978, p 213

Braunwald, E, *et al*: Mechanism of Contractility of the Normal and Failing Heart. Boston, Little, Brown, 1968

Chatterjee, K, and Parmley, W W Vasodilator therapy for chronic heart failure *Annu Rev Pharmacol Toxicol*, *20* 475, 1980

Corday, E, and Swan, H J C (ed) Clinical Strategies in Ischemic Heart Disease Baltimore Williams & Wilkins, 1979

Fein, F S, and Sonnenblick, E. H Diabetic cardiomyopathy *Prog. Cardiovasc Dis*, *27* 255, 1985

Francis, G S Neurohumoral mechanisms involved in congestive heart failure *Am J. Cardiol*, *55* 15A, 1985

Cibba, C L Cardiac energetics *Physiol Rev*, *58* 174, 1978

Gilbert I and Henning, R J Adenocarcinoma of the lung presenting with pericardial tamponade Report of a case and review of the literature *Heart Lung*, *14* 83, 1985

Guyton A C The systemic venous system in cardiac failure *J Chronic Dis*, *9* 465, 1959

Guyton, A C, *et al* Cardiac Output and Its Regulation Philadelphia, W B Saunders Co, 1973

Kirk, E S, *et al* Problems in the cardiac performance regulation of coronary blood flow and the physiology of heart failure *In* MTP International Review of Science Physiology Vol 1 Baltimore, University Park Press, 1974, p 229

Kones, R. J Cardiogenic shock Therapeutic implications of altered myocardial energy balance *Angiology*, *25* 317, 1974

Kones, R J Cardiogenic Shock Mechanisms and Management New York, Future Publishing Co, 1975

Lange, L G, and Sobel, B E Pharmacological salvage of myocardium *Annu Rev Pharmacol Toxicol*, *22* 115, 1982

McManus Cardiovascular Complications of Exercise Chicago, Year Book Medical Publishers, 1985

Norris, R M Myocardial Infarction Its Presentation, Pathogenesis and Treatment New York, Churchill Livingstone, Inc, 1982

Orrego, F Calcium and the mechanism of action of digitalis. *Gen Pharmacol*, *15* 273, 1984

Rapaport, E (ed) Current Controversies in Cardiovascular Disease Philadelphia, W B Saunders Co, 1980

Regan, T J Congestive heart failure in the diabetic *Annu Rev Med*, *34* 161, 1983

Samuel, P, *et al* The role of diet in the etiology and treatment of atherosclerosis *Annu Rev Med*, *34* 179, 1983

Silver, M D Cardiovascular Pathology New York, Churchill Livingstone, Inc, 1983

Stallones, R. A Ischemic heart disease and lipids in blood and diet. *Annu Rev Nutr*, *3* 155, 1983

Stone, H L, *et al* Progressive changes in cardiovascular function after unilateral heart irradiation. *Am J Physiol*, *206* 289, 1964

Stone, H L, *et al* Ventricular function following radiation damage of the right ventricle *Am J Physiol*, *211* 1209, 1966.

Sullivan, J M Drug suppression of the angiotensin system in congestive heart failure *Annu Rev Med*, *34* 169, 1983

Svennerholm, L, *et al* (eds) Structure and Function of Gangliosides. New York, Plenum Press, 1979

Swan, H J C, and Parmley, W W Congestive heart failure *In* Sodeman, W A, Jr, and Sodeman, T M (eds) Pathologic Physiology. Mechanisms of Disease, 6th Ed. Philadelphia, W B Saunders Co, 1979, p 313

Weisfeldt, M L The aging heart. *Hosp Pract (Off)*, *20* 115, 125, 1985

Wenger, N K (ed) Exercise and the Heart. Philadelphia, F. A Davis Co, 1978.

Willis, J. (ed) The Heart. *Update New York, McGraw Hill*, 1979

Winkelman, J L Essentials of Basic Life Support. Minneapolis, Burgess, 1981.

Winwood, R. S. Essentials of Clinical Diagnosis in Cardiology Baltimore, E Arnold, 1981.

Zelis, R, *et al* · Peripheral circulatory control mechanisms in congestive heart failure *Am J Cardiol*, *32* 481, 1973

Zelis, R., *et al*.. Cardiocirculatory dynamics in the normal and failing heart. *Annu Rev. Physiol*, *43* 455, 1981

27

Heart Sounds; Dynamics of Valvular and Congenital Heart Defects

THE HEART SOUNDS

The function of the heart valves was discussed in Chapter 13, and it was pointed out that closure of the valves is associated with audible sounds On the other hand, no sounds occur when the valves open The purpose of the present section is to discuss the factors that cause the sounds in the heart, under both normal and abnormal conditions

NORMAL HEART SOUNDS

Listening with a stethoscope to a normal heart, one hears a sound usually described as "lub, dub, lub, dub ." The "lub" is associated with closure of the A-V valves at the beginning of systole and the "dub" with closure of the semilunar valves at the end of systole The "lub" sound is called the *first heart sound* and the "dub" the *second heart sound* because the normal cycle of the heart is considered to start with the beginning of systole

Causes of the First and Second Heart Sounds. Closure of the valves in any pump system usually causes a certain amount of noise because the valves close solidly and suddenly over some opening, setting up vibrations in the fluid or walls of the pump In the heart, the valves are cushioned by blood, so that it is difficult to understand why these valves create as much sound as they do

The earliest suggestion for the cause of the heart sounds was that the slapping together of the valve leaflets themselves sets up vibrations, but this has now been shown to cause little if any of the sound because of the cushioning effect of the blood Instead, the cause is *vibration of the taut valves immediately after closure*, as well as *vibration of the adjacent blood, walls of the heart, and major vessels around the heart* That is, in the generation of the first heart sound, contraction of the ventricles causes sudden backflow of blood against A-V valves, causing them to bulge toward the atria the chordae tendineae abruptly stop the backbulging The elastic tautness of the valves then causes the backsurging blood to bounce forward again into each respective ventricle This sets the blood and the ventricles as well as the valves into vibration and also causes vibrating turbulence in the blood The vibrations then travel to the chest wall where they can be heard as sound by the stethoscope.

The second heart sound results from vibration of the taut, suddenly closed semilunar valves and from vibration of the blood and the walls of the pulmonary artery, the aorta, and, to a much lesser extent, the ventricles. When the semilunar valves close, they bulge backward toward the ventricles, and their elastic stretch recoils the blood back into the arteries, which causes a short period of reverberation of blood back and forth between the walls of the arteries and the valves and also between the valves and the ventricular walls. The vibrations set up in the arterial walls are then transmitted along the arteries at the velocity of the pulse wave. When the vibrations of the vessels or ventricles come into contact with a "sounding board," such as the chest wall, they create sound that can be heard.

Durations and Frequencies of the First and Second Heart Sounds. The duration of each of the heart sounds is slightly more than 0.1 second, the first sound lasts about 0 14 second and the second about 0 11 second Both of them are described as very low-pitched sounds, the first lower than the second

Sound consists of vibrations of different frequencies Figure 27–1 illustrates by the shaded area the amplitudes of the different frequencies in the heart sounds and murmurs, showing that these are composed of frequencies ranging all the way from a few cycles per second to more than 1000 cycles per second, with the maximum amplitude of vibration occurring at a frequency of about 24 cycles per second, which is actually below the audible range of the ear

Also shown in Figure 27–1 is a curve called the "threshold of audibility," which depicts the capability of the ear to hear sounds of different amplitudes Note that in the very low frequency range the heart vibrations have a high amplitude, but the threshold of audibility is so high that ordinarily the heart vibrations below approximately 30 to 50 cycles per second are not heard by the ears Then above about 500 cycles per second, the heart sounds are so weak that, despite a low threshold of audibility, no frequencies in this range are heard For practical purposes, then, we can consider that all

Emphasis

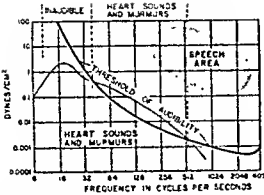


Figure 27–1. Amplitude of different frequency vibrations in the heart sounds and heart murmurs in relation to the threshold of audibility, showing that the range of sounds that can be heard is between about 40 and 500 cycles per second (Modified from Butterworth, Chassin, and McGrath Cardiac Auscultation Grune and Stratton.)

the *audible* heart sounds lie in the range of approximately 40 to 500 cycles per second despite the fact that the maximum amplitude of vibration occurs at the very low frequency of 24 cycles per second

Both the first and second heart sounds have a mixture of frequencies in the entire audible range of the heart sounds, though the first sound has slightly more low frequency sound than does the second heart sound

The reason the frequency of the first heart sound is lower than that of the second sound is probably twofold First, the *elastic modulus* of the A-V valves and of the walls of the ventricles is far less than the elastic modulus of the semilunar valves and the arterial walls It is well known that any mechanical vibrating system having a low elastic modulus oscillates at a lower frequency than a system having a greater modulus Second, the *mass of blood* in the ventricles is much greater than that in the initial parts of the great vessels, which means that the inertia of the vibrating mass is also much greater This also would cause the first heart sound to have a lower frequency than the second

The reason the second heart sound is shorter in duration than the first is probably that the second sound is "damped" out by the vascular walls much more rapidly than is the first heart sound by the ventricular walls.

Loudness of the First and Second Heart Sounds. The loudness of the heart sounds is almost directly proportional to the *rate of change* of the pressure differences across the valves For instance, when the onset of systole is very rapid, the intraventricular pressure rises very rapidly during the isometric period of ventricular contraction; yet the pressure on the opposite side of the A-V valves remains only a few millimeters Hg. Thus, the rate of change of the pressure difference between the two sides of the valves is very great, and as a result the first heart sound is loud Also, when the heart is very active, such as during and immediately following exercise, the force of contraction of the ventricle is greatly enhanced, so that the first heart sound is in this instance also greatly accentuated. Conversely, in a weakened heart in which the onset of contraction is sluggish, the loudness of the first sound is greatly diminished.

In the case of the second heart sound, it is the rate of decrease in ventricular pressure at the end of systole that determines the loudness. The magnitude of this rate of decrease is determined mainly by the level of ventricular systolic pressure at the time the valve closes. In a person who has hypertension, the pressure at the time the aortic valve closes may be as great as 200 mm Hg, so that the intraventricular pressure falls twice as rapidly as normally, all the way to zero in a few hundredths of a second. Therefore, the aortic sound is markedly accentuated Likewise, in pulmonary hypertension the pulmonic sound is greatly accentuated. On the other hand, when the arterial pressure is low, such as in shock or in the terminal stages of cardiac failure, the second heart sound is diminished to a very low intensity

The Third Heart Sound. Occasionally a third heart sound is heard at the beginning of the middle third of diastole A logical, but yet unproved, explanation of this sound is oscillation of blood back and forth between the walls of the ventricles initiated by inrushing blood from the atria. This is analogous to running water from a faucet into a sack, the inrushing water reverberating back and forth between the walls of the sack to cause vibrations in the walls.

The third heart sound is an extremely weak rumble of such low frequency that it usually cannot be heard with a stethoscope, but it can be recorded frequently in the phonocardiogram. The very low frequency of this sound presumably results from the flaccid, inelastic condition of the heart during diastole Also, the reason the third heart sound does not occur until the middle third of diastole is presumably that in the early part of diastole the heart is not filled sufficiently to create even the small amount of elastic tension in the ventricles necessary for reverberation. The reason the third heart sound does not continue into the latter part of diastole is presumably that little blood flows into the ventricles during the latter part of diastole, so that no initiating stimulus then exists for causing reverberation.

The Atrial Heart Sound (Fourth Heart Sound). An atrial heart sound can be recorded in many persons in the phonocardiogram, but it can almost never be heard with a stethoscope because of its low frequency—usually 20 cycles per second or less. This sound occurs when the atria contract, and presumably it is caused by inrush of blood into the ventricles, which initiates vibrations similar to those of the third heart sound.

AREAS FOR AUSCULTATION OF NORMAL HEART SOUNDS

Listening to the sounds of the body, usually with the aid of a stethoscope, is called *auscultation*. Figure 27–2 illustrates the areas of the chest wall from which the different valvular sounds can best be distinguished. With the stethoscope placed in any one of the special valvular areas, the sounds from all the other valves can also be heard, though the sound from the special valve is as loud, *relative to the other sounds*, as it ever will be. The cardiologist distinguishes the sounds from the different valves by a process of elimination; that is, he moves the stethoscope from one area to another, noting the loudness of the sounds in different areas and gradually picking out the sound components from each valve.

The areas for listening to the different heart sounds are not directly over the valves themselves. The aortic area is upward along the aorta, the pulmonic area is upward along the pulmonary artery, the tricuspid area is over the right ventricle, and the mitral area is over

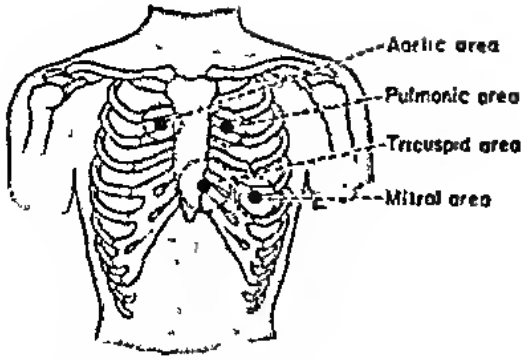


Figure 27–2. Chest areas from which each valve sound is best heard

the apex of the heart, which is the portion of the left ventricle nearest the surface of the chest because the heart is rotated so that most of the left ventricle lies behind the right ventricle In other words, the sounds caused by the A-V valves are transmitted to the chest wall through each respective ventricle, and the sounds from the semilunar valves are transmitted especially along the great vessels leading from the heart. This transmission of sounds is in keeping with the concept that vibrations in the ventricles or large arteries are the cause of the heart sounds

THE PHONOCARDIOGRAM

If a microphone specially designed to detect low frequency sound waves is placed on the chest, the heart sounds can be amplified Recording is possible by a high-speed recording apparatus, such as an oscilloscope or a high-speed pen recorder, these were described in Chapters 10 and 15 for recording nerve potentials and electrocardiograms The recording is called a *phonocardiogram*, and the heart sounds appear as waves, as illustrated schematically in Figure 27–3. Record A is a recording of normal heart sounds, showing the vibrations of the first, second, and third heart sounds and even the atrial sound Note specifically that the third and atrial heart sounds are each a very low rumble The third heart sound can be recorded in only one third to one half of all persons, and the atrial heart sound can be recorded in perhaps one fourth of all persons

VALVULAR LESIONS

Rheumatic Valvular Lesions. By far the greatest number of valvular lesions results from rheumatic fever. Rheumatic fever is an autoimmune or allergic disease in which the heart valves are likely to be damaged or destroyed It is initiated by streptococcal toxin in the following manner

The entire sequence of events almost always begins with a preliminary streptococcal infection caused specifically by group A hemolytic streptococci, such as a sore throat, scarlet fever, or middle ear infection The streptococci release several different proteins against which antibodies are formed, the most important of which seems to be a protein called the "M" antigen The [illegible]bodies then react with many different tissues of the [illegible], causing either immunologic or allergic damage [illegible] reactions continue to take place as long as the [illegible]tibodies persist in the blood—six months or more As a result, rheumatic fever causes damage in many parts of the body but especially in certain very susceptible areas such as the heart valves. The degree of heart valve damage is directly correlated with the titer and the persistence of these antibodies. Principles of immunity relating to this type of reaction were discussed in Chapter 6, and it is also noted in Chapter 38 that acute glomerular nephritis has a similar basis.

In rheumatic fever, large hemorrhagic, fibrinous, bulbous lesions grow along the inflamed edges of the heart valves Because the mitral valve receives more trauma during valvular action than any of the other valves, this valve is the one most often seriously damaged, and the aortic valve is second most frequently damaged The tricuspid and pulmonary valves are also often involved, but much less severely, probably because the stresses acting on these valves are slight compared with those in the left ventricle

Scarring of the Valves. The lesions of acute rheumatic fever frequently occur on adjacent valve leaflets simultaneously so that the edges of the leaflets become stuck together Then, weeks, months, or years later, the lesions become scar tissue, permanently fusing portions of the leaflets Also, the free edges of the leaflets, which are normally filmy and free-flapping, become solid, scarred masses

A valve in which the leaflets adhere to each other so extensively that blood cannot flow through satisfactorily is said to be *stenosed* On the other hand, when the valve edges are so destroyed by scar tissue that they cannot close when the ventricles contract, *regurgitation*, or backflow, of blood occurs when the valve should be closed Stenosis usually does not occur without the coexistence of at least some degree of regurgitation, and vice versa Therefore, when a person is said to have stenosis or regurgitation, it is usually meant that one predominates over the other

Other Causes of Valvular Lesions. Stenosis or lack of one or more leaflets of a valve frequently occurs as a congenital defect Complete lack of leaflets is rare, though stenosis is common, as is discussed later in this chapter

ABNORMAL HEART SOUNDS CAUSED BY VALVULAR LESIONS

As illustrated by the phonocardiograms of Figure 27–3, many abnormal heart sounds, known as "mur-

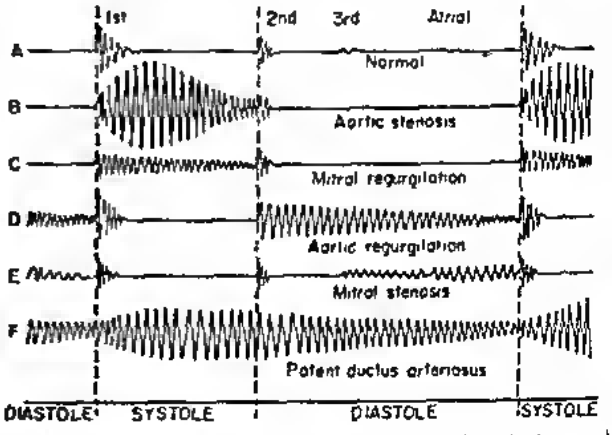


Figure 27–3. Phonocardiograms from normal and abnormal hearts.

murs," occur when there are abnormalities of the valves, as follows:

The Murmur of Aortic Stenosis. In aortic stenosis, blood is ejected from the left ventricle through only a small opening of the aortic valve. Because of the resistance to ejection, the pressure in the left ventricle rises sometimes to as high as 300 mmHg while the pressure in the aorta is still normal. Thus, a nozzle effect is created *during systole*, with blood jetting at tremendous velocity through the small opening of the valve. This causes *severe turbulence* of the blood in the root of the aorta. The turbulent blood impinging against the aortic walls causes intense vibration, and a loud murmur is transmitted throughout the upper aorta and even into the larger arteries of the neck. This sound is harsh and in severe stenosis occasionally is so loud that it can be heard several feet away from the patient. Also, the sound vibrations can often be felt with the hand on the upper chest and lower neck, a phenomenon known as a "thrill."

The Murmur of Aortic Regurgitation. In aortic regurgitation no sound is heard during systole, but *during diastole* blood flows backward from the aorta into the left ventricle, causing a "blowing" murmur of relatively high pitch and with a swishing quality heard maximally over the left ventricle. This murmur results from *turbulence* of blood jetting backward into the blood already in the left ventricle.

The sound of aortic regurgitation is not nearly so loud as that of aortic stenosis, mainly because the pressure differential between the aorta and left ventricle in regurgitation is not so great as it usually is in stenosis.

If the aortic valve is so badly destroyed that essentially all the return of blood from the arterial tree into the heart takes place during the first portion of diastole, the sound of aortic regurgitation may not be heard at all during the latter portion of diastole. Therefore, aortic regurgitation murmurs lasting throughout diastole sometimes mean less severely damaged valves than weaker murmurs heard only during the early part.

The Murmur of Mitral Regurgitation. In mitral regurgitation blood flows backward through the mitral valve *during systole*. This also causes a high frequency "blowing," swishing sound, which is transmitted most strongly into the left atrium. However, the left atrium is so deep within the chest that it is difficult to hear this sound directly over the atrium. As a result, the sound of mitral regurgitation is transmitted to the chest wall mainly through the left ventricle, and it is usually heard best at the apex of the heart.

The blowing, swishing quality of the mitral regurgitation murmur, like that of aortic regurgitation, is presumably caused by the turbulence of blood jetting backward through the mitral valve into the blood already in the left atrium or against the atrial wall. The quality of mitral regurgitation murmur is almost the same as that of aortic regurgitation, but it occurs during systole rather than diastole.

The Murmur of Mitral Stenosis. In mitral stenosis, blood passes with difficulty from the left atrium into the left ventricle and, because the pressure in the left atrium rarely rises above 35 mm Hg except for short periods of time, a great pressure differential forcing blood from the left atrium into the left ventricle never develops. Consequently, the abnormal sounds heard in mitral stenosis are usually weak.

During the early part of diastole, the ventricle has so little blood in it and its walls are so flabby that blood does not reverberate back and forth between the walls of the ventricle. For this reason, even in severe mitral stenosis, no murmur at all might be heard during the first third of diastole. However, after the first third the ventricle is stretched enough for blood to reverberate, and a low rumbling murmur then often begins. This murmur is of such low pitch that it is difficult to hear, but, with the aid of a proper stethoscope (the "bell" type), one can usually discern very low frequency sounds of 30 to 50 cycles per second. In mild stenosis the murmur lasts only during the first half of the middle third of diastole, but in severe stenosis it can begin early in diastole and persist for the whole remainder of diastole. On the other hand, one can often feel low frequency vibrations, called a "thrill," over the apex of the heart despite the fact that the sound itself may be weak. *The reason for this is that the frequency of* vibration in the ventricle is often so low that it cannot be heard but yet can be felt.

Often in early stages of mitral stenosis a *presystolic murmur* may be heard. This presystolic murmur is caused by the momentarily increased left atrial pressure resulting from atrial contraction.

Phonocardiograms of Valvular Murmurs. Phonocardiograms B, C, D, and E of Figure 27–3 illustrate, respectively, idealized records obtained from patients with aortic stenosis, mitral regurgitation, aortic regurgitation, and mitral stenosis. It is obvious from these phonocardiograms that the aortic stenotic lesion causes the loudest of all these murmurs, and the mitral stenotic lesion causes the weakest; a murmur of very low frequency and rumbling quality. The phonocardiograms show how the intensity of the murmurs varies during different portions of systole and diastole, and the relative timing of each murmur is also evident. Note especially that the murmurs of aortic stenosis and mitral regurgitation occur only during systole, while the murmurs of aortic regurgitation and mitral stenosis occur only during diastole. If a student does not understand this timing, a moment's pause should be taken until it is understood.

ABNORMAL CIRCULATORY DYNAMICS IN VALVULAR HEART DISEASE

DYNAMICS OF THE CIRCULATION IN AORTIC STENOSIS AND AORTIC REGURGITATION

In *aortic stenosis*, the left ventricle fails to empty adequately, whereas in *aortic regurgitation* blood returns to the ventricle after the ventricle has been emptied. Therefore, in either case, the *net stroke volume output* of the heart is reduced.

However, several important compensations take place that can ameliorate the severity of the circulatory defects. Some of these are the following:

Hypertrophy of the Left Ventricle. In both aortic stenosis and regurgitation, the left ventricular musculature hypertrophies, and in regurgitation the ventricle also enlarges to hold all the regurgitant blood. Sometimes the left ventricle muscle mass increases as much as four- to fivefold, creating a

heart When the aortic valve is seriously stenosed, this hypertrophied muscle allows the left ventricle to develop as much as 400 mm Hg intraventricular pressure during occasional periods of peak activity, even at rest the pressure differential across the stenotic valve is sometimes 100 mm Hg. In severe aortic regurgitation, the hypertrophied muscle allows the left ventricle sometimes to pump a stroke volume output as high as 250 ml, though as much as three fourths of this blood on occasion returns to the ventricle during diastole

Increase in Blood Volume. Another effect that helps to compensate for the diminished net pumping by the left ventricle is increased blood volume This results from an initial slight decrease in arterial pressure plus peripheral circulatory reflexes that this decrease induces, both of which diminish renal output of urine until the blood volume increases and the mean pressure returns to normal Also, red cell mass eventually increases because of a slight degree of tissue hypoxia.

The increase in blood volume tends to increase venous return to the heart This in turns increases the ventricular end-diastolic volume, causing the left ventricle to pump its very high pressure in aortic stenosis or its very high stroke volume output in aortic regurgitation

Eventual Failure of the Left Ventricle, and Development of Pulmonary Edema

In the early stages of aortic stenosis or aortic regurgitation, the intrinsic ability of the left ventricle to adapt to increasing loads prevents significant abnormalities in circulatory function in the person at rest other than increased work output required of the left ventricle Therefore, marked degrees of aortic stenosis or aortic regurgitation often occur before the person knows that he has serious heart disease, such as resting left ventricular systolic pressures as high as 200 mm Hg in aortic stenosis, or left ventricular stroke volume outputs as high as double normal in aortic regurgitation

However, beyond critical stages of development of these two aortic lesions, the left ventricle finally cannot keep up with the work demand, and as a consequence the left ventricle dilates and cardiac output begins to fall while blood simultaneously dams up in the left atrium and lungs behind the failing left ventricle The left atrial pressure rises progressively, and at pressures

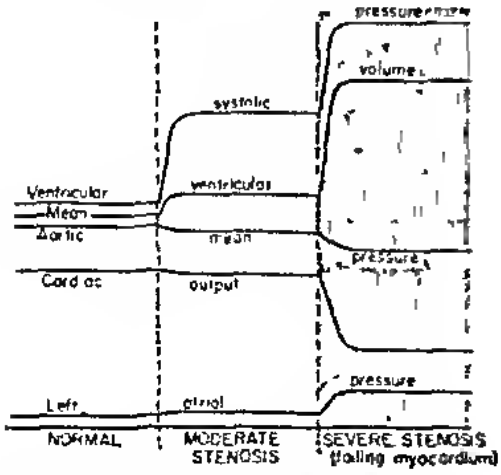


Figure 27–4. Circulatory dynamics in aortic stenosis.

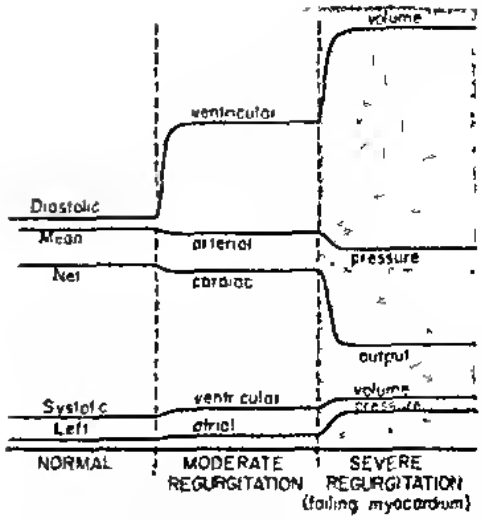


Figure 27–5. Circulatory dynamics in aortic regurgitation.

above 30 to 40 mm Hg mean atrial pressure edema appears in the lungs, as is discussed in detail in Chapter 31

The progressive changes in circulatory dynamics are diagrammed in Figures 27–4 and 27–5 for aortic stenosis and regurgitation, respectively

Myocardial Ischemia in Aortic Valvular Disease. Because of the very high intraventricular pressure during systole in aortic stenosis, very little blood flows through the coronary arteries during systole. Therefore, extra blood flow is required during diastole to make up the difference Unfortunately, the hypertrophied muscle of the left ventricle also frequently has a relatively deficient coronary vasculature Finally, the intraventricular pressure sometimes remains high during diastole, thereby compressing the inner layers of the heart and also diminishing coronary blood flow. For all these reasons, the patient frequently experiences severe degrees of coronary ischemia and angina.

In aortic regurgitation the problem is further compounded because the diastolic aortic pressure frequently falls very low as the aortic blood regurgitates back into the ventricle Since most of the left ventricular coronary blood flow occurs during diastole, this low pressure can be particularly detrimental to the flow The effect is especially serious for the subendocardial myocardium in which almost zero flow occurs during systole Therefore, again, coronary ischemia occurs, with concomitant anginal pain

DYNAMICS OF MITRAL STENOSIS AND MITRAL REGURGITATION

In mitral stenosis blood flow from the left atrium into the left ventricle is impeded, and in mitral regurgitation much of the blood that has flowed into the left ventricle leaks back into the left atrium during systole rather than being pumped into the aorta. Therefore, the effect is reduced net movement of blood from the left atrium into the left ventricle.

Pulmonary Edema in Mitral Valvular Disease. Obviously the buildup of blood in the left atrium causes progressive increase in left atrial pressure, and this can result eventually in the development of serious pulmonary edema. Ordinarily, lethal edema will not occur

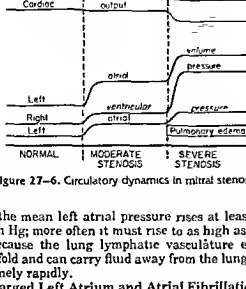


Figure 27–6. Circulatory dynamics in mitral stenosis

until the mean left atrial pressure rises at least above 30 mm Hg; more often it must rise to as high as 40 mm Hg because the lung lymphatic vasculature enlarges manyfold and can carry fluid away from the lung tissues extremely rapidly.

Enlarged Left Atrium and Atrial Fibrillation. The high left atrial pressure also causes progressive enlargement of the left atrium, which increases the distance that the cardiac impulse must travel in the atrial wall. Eventually, this pathway becomes so long that it predisposes to the development of circus movements Therefore, in late stages of mitral valvular disease, especially mitral stenosis, atrial fibrillation usually occurs This further reduces the pumping effectiveness of the heart and, therefore, causes still further cardiac debility

Compensations in Mitral Valvular Disease. As also occurs in aortic valvular disease and in many types of congenital heart disease, the blood volume increases in mitral valvular disease, principally because of diminished blood flow to the kidneys This increases venous return to the heart, thereby helping to overcome the effect of the cardiac debility to reduce cardiac output Therefore, cardiac output does not fall more than minimally until the late stages of mitral valvular disease.

As the left atrial and pulmonary capillary pressures rise, blood also begins to dam up in the pulmonary artery, and the incipient edema of the lungs causes intense pulmonary arteriolar constriction, these two effects together then increase pulmonary arterial pressure sometimes to as high as 60 mm Hg This, in turn, causes hypertrophy of the right heart, which partially compensates for its increased work load

Figure 27–6 illustrates the progressive changes in circulatory dynamics in mitral stenosis

Essentially the same effects occur in mitral regurgitation. Indeed, these two conditions more often than not occur together in varying degrees

CIRCULATORY DYNAMICS DURING EXERCISE IN PATIENTS WITH VALVULAR LESIONS

During exercise, very large quantities of venous blood are returned to the heart from the peripheral circulation Therefore, all of the dynamic abnormalities that occur in the different types of valvular heart disease become tremendously exacerbated. Even in mild valvular heart disease, in which the symptoms may be completely unrecognizable at rest, severe symptoms often develop during heavy exercise. For instance, in patients with aortic valvular lesions, exercise can cause acute left ventricular failure followed by *acute pulmonary edema.* Also, in patients with mitral disease, exercise can cause so much damming of blood in the lungs that serious pulmonary edema ensues within minutes.

Even in the mildest cases of valvular disease, the patient finds that his *cardiac reserve* is diminished in proportion to the severity of the valvular dysfunction. That is, the cardiac output does not increase as it should during exercise. Therefore, the muscles of the body fatigue rapidly.

ABNORMAL CIRCULATORY DYNAMICS IN CONGENITAL HEART DEFECTS

Occasionally, the heart or its associated blood vessels are malformed during fetal life; the defect is called a *congenital anomaly* Basically, there are three major types of congenital anomalies of the heart and its associated vessels. (1) *stenosis* of the channel of blood flow at some point in the heart or in a closely allied major vessel, (2) an abnormality that allows blood to flow directly from the left heart or aorta to the right heart or pulmonary artery, thus by-passing the systemic circulation—this is called a *left-to-right shunt*, and (3) an abnormality that allows blood to flow from the right heart or pulmonary artery directly into the left heart or aorta, thus bypassing the lungs—this is called a *right-to-left shunt*

Stenotic Lesions. The three most common types of stenotic lesions are *coarctation of the aorta, congenital pulmonary stenosis,* and *congenital aortic stenosis.* The effect of coarctation of the aorta on the circulation in relation to the regulation of arterial pressure was discussed in Chapter 22, and congenital aortic stenosis causes almost exactly the same alterations in circulatory dynamics as occur in rheumatic aortic valvular stenosis. The dynamics of pulmonary stenosis will be discussed later in this chapter.

Left-to-Right Shunts. In all left-to-right shunts, blood is shunted directly from the left heart back to the right heart without going through the systemic circulation Compensatory effects then occur in the circulatory system to increase the total cardiac output, thereby making up for most of the extra blood flow through the shunt. For instance, if 50 per cent of all the blood pumped by the left heart is shunted directly back to the right heart, then the compensatory mechanisms normally increase the cardiac output to almost double normal so that the amount of blood flowing through the systemic vessels is still almost normal despite the shunt. However, this causes an extra load on the heart and therefore usually causes the heart to fail at an early age. Also, it reduces the cardiac reserve of the person throughout his life. Some of the types of left-to-right shunts are *patent ductus arteriosus, interatrial septal defect, interventricular septal defect,* and a *direct communication between the aorta and pulmonary artery at their bases.*

Right-to-Left Shunts. In all right-to-left shunts, much of the blood bypasses the lungs and therefore fails to become oxygenated. Consequently, large quantities of venous blood enter the arterial system directly, and the patient is usually cyanotic (blue) all the time. The

most common type of right-to-left shunt is the *tetralogy of Fallot*, which is described later; several other right-to-left shunts are some variation of this abnormality

Causes of Congenital Anomalies. One of the most common causes of congenital heart defects is a virus infection in the mother during the first trimester of pregnancy when the fetal heart is being formed. Defects are particularly prone to develop when the mother contracts German measles at this time—so often indeed that obstetricians advise termination of pregnancy if German measles occurs in the first trimester However, some congenital defects of the heart are hereditary, because the same defect has been known to occur in identical twins and also in succeeding generations Children of patients surgically treated for congenital heart disease have ten times as much chance of having congenital heart disease as do other children Congenital defects of the heart are frequently also associated with other congenital defects of the body

PATENT DUCTUS ARTERIOSUS—A LEFT-TO-RIGHT SHUNT

During fetal life the lungs are collapsed, and the elastic factors that keep the alveoli collapsed also keep the blood vessels collapsed Therefore, in the collapsed lung the resistance to blood flow is perhaps five times as great as it is in the inflated lung. For this reason, the pulmonary arterial pressure is high in the fetus. On the other hand, because of very low resistance through the large vessels of the placenta the pressure in the aorta is lower than in the pulmonary artery, causing almost all the pulmonary arterial blood to flow through the ductus arteriosus into the aorta rather than through the lungs This allows immediate recirculation of the blood through the systemic arteries of the fetus Obviously, this lack of blood flow through the lungs is not detrimental to the fetus because the blood is oxygenated by the placenta of the mother

Closure of the Ductus. As soon as the baby is born its lungs inflate, and not only do the alveoli fill with air but also the resistance to blood flow through the pulmonary vascular tree decreases tremendously, allowing pulmonary arterial pressure to fall. Simultaneously, the aortic pressure rises because of sudden cessation of blood flow through the placenta Thus, the pressure in the pulmonary artery falls, while that in the aorta rises. As a result, forward blood flow through the ductus ceases suddenly at birth, and blood begins to flow backward from the aorta into the pulmonary artery. This new state of blood flow causes the ductus arteriosus to become occluded within a few hours to a few days in most babies so that blood flow through the ductus does not persist. The ductus probably closes because the aortic blood now flowing through the ductus has about two times as high an oxygen concentration as the pulmonary artery blood that had been flowing in the ductus, and the oxygen constricts the ductus muscle Reduction of prostaglandins acting on the ductus seems to help in this closure as well This will be discussed further in Chapter 83. In some instances it takes several weeks for the ductus to close completely, and in about 1 of every 5500 babies the ductus never closes, causing the condition known as *patent ductus arteriosus*, which is illustrated in Figure 27–7.

Dynamics of Persistent Patent Ductus. During the early months of an infant's life a patent ductus usually does not cause severely abnormal dynamics because the blood pressure of the aorta then is not much higher than the pressure in the pulmonary artery, and only a small amount of blood flows backward into the pulmonary system However, as the child grows older the differential between the pressure in the aorta and that in the pulmonary artery progressively increases, with corresponding increase in the backward flow of blood from the aorta to the pulmonary artery Also, the diameter of the partially closed ductus often increases with time, making the condition worse

Recirculation Through the Lungs. In the older child with a patent ductus, as much as half to two thirds of the blood flows into the pulmonary artery, then through the lungs, into the left atrium, and finally back into the left ventricle, passing through this lung and left heart circuit two or more times for every one time that it passes through the systemic circulation.

These persons do not show cyanosis until the heart fails or until the lungs become congested. Indeed, the arterial blood is often better oxygenated than normally because of the extra times of passage through the lungs Furthermore, in the early stages of patent ductus arteriosus the quantity of blood flowing into the systemic aorta remains essentially normal because the quantity of blood returning to the heart from the peripheral circulatory system is normal Yet, because of the tremendous accessory flow of blood around and around

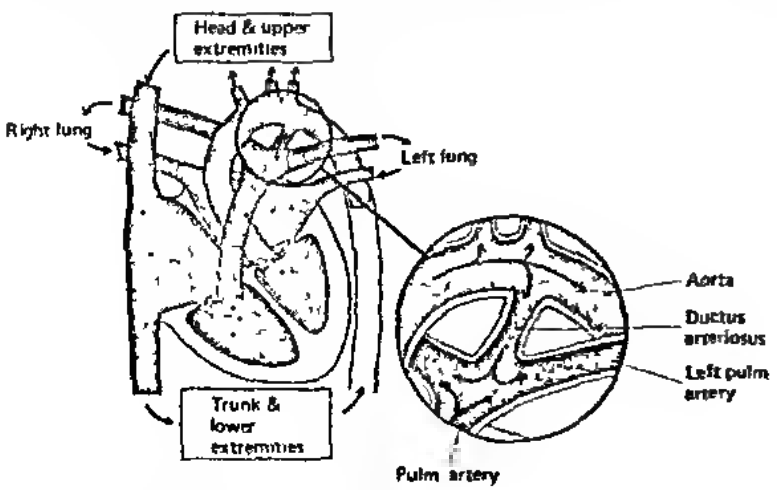


Figure 27–7. Patent ductus arteriosus, illustrating the degree of blood oxygenation in the different parts of the circulation.

through the lungs and left side of the heart, the output of the left ventricle in patent ductus arteriosus is often two to three times normal.

Diminished Cardiac and Respiratory Reserve. The major effects of patent ductus arteriosus on the patient are low cardiac and respiratory reserve The left ventricle is already pumping approximately two or more times the normal cardiac output, and the maximum that it can possibly pump is about four to six times normal Therefore, during exercise the net blood flow through the remainder of the body can be increased much less than usual. Under basal conditions, patients usually appear normal except for possible heaving of the chest with each beat of the heart, but with even moderately strenuous exercise they are likely to become weak and occasionally even faint from momentary heart failure Also, the high pressures in the pulmonary vessels often lead to pulmonary congestion

As a result of the increased load on the heart and especially because the pulmonary congestion becomes progressively more severe with age, most patients with uncorrected patent ductus die between the ages of 20 and 40 years.

The Machinery Murmur. In the infant with patent ductus arteriosus, occasionally no abnormal heart sounds are heard because the quantity of reversed blood flow may be insufficient As the baby grows older, reaching the age of one to three years, a harsh, blowing murmur begins to be heard in the pulmonic area of the chest. This sound is much more intense during systole when the aortic pressure is high and much less intense during diastole when the aortic pressure falls very low, so that the murmur waxes and wanes with each beat of the heart, creating the so-called "machinery murmur" The idealized phonocardiogram of this murmur is shown in Figure 27–3F.

Surgical Treatment. Surgical treatment of patent ductus arteriosus is extremely simple, for all one needs to do is to ligate the patent ductus or to divide it and sew the two ends.

INTERVENTRICULAR SEPTAL DEFECT—A LEFT-TO-RIGHT SHUNT

Because the systolic pressure in the left ventricle is normally about five times that in the right ventricle, a large amount of blood flows from the left to the right ventricle whenever a hole occurs in the septum, called an *interventricular defect* The excess flow of blood into the right ventricle in turn increases the pressure in the right ventricle. As a result, the right ventricle hypertrophies, sometimes to such an extent that its muscular wall approximately equals that of the left ventricle.

Diagnosis of an interventricular septal defect is based on (1) the presence of a systolic blowing murmur heard over the anterior projection of the heart, (2) high right ventricular systolic pressure recorded from a catheter, and (3) the presence of oxygenated blood in a blood *sample removed through a catheter in the right ventricle*, this blood having leaked backward from the left ventricle

Blood flowing from the left ventricle into the right ventricle passes one or more times through the lungs and then back to the left ventricle again before finally entering the peripheral circulatory system This condition, therefore, is analogous to patent ductus arteriosus except that the flow of blood from the systemic circulation to the pulmonary circulation occurs only during systole rather than during both systole and diastole as in patent ductus.

The septal defect can be treated surgically by placing a patch over the defect.

TETRALOGY OF FALLOT—A RIGHT-TO-LEFT SHUNT

Tetralogy of Fallot is illustrated in Figure 27–8, in which it will be noted that four different abnormalities of the heart occur simultaneously

First, the aorta originates from the right ventricle rather than the left, or it overrides the septum as shown in the figure, receiving blood from both ventricles

Second, the pulmonary artery is stenosed so that much less than normal amounts of blood pass from the right ventricle into the lungs, instead the blood passes into the aorta

Third, blood from the left ventricle flows through a ventricular septal defect into the right ventricle and then into the aorta or directly into the overriding aorta

Fourth, because the right side of the heart must pump large quantities of blood against the high pressure in the aorta, its musculature is highly developed, causing an enlarged right ventricle

Abnormal Dynamics. It is readily apparent that the major physiological difficulty caused by tetralogy of Fallot is the shunting of blood past the lungs without its becoming oxygenated. As much as 75 per cent of the venous blood returning to the heart may pass directly from the right ventricle into the aorta without becoming oxygenated Tetralogy of Fallot is the major cause of cyanosis in babies ("blue babies")

A diagnosis of tetralogy of Fallot is usually based on (1) the fact that the baby is blue, (2) records of high systolic pressure in the right ventricle recorded through

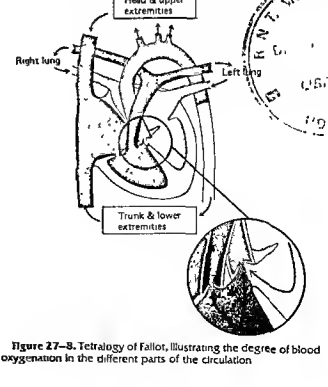


Figure 27–8. Tetralogy of Fallot, illustrating the degree of blood oxygenation in the different parts of the circulation

a catheter, (3) characteristic changes in the radiologic silhouette of the heart showing an enlarged right ventricle, and (4) angiograms showing abnormal blood flow through the interventricular septal defect, the overriding aorta, and the pulmonary stenotic area

Surgical Treatment. In recent years tetralogy of Fallot has been treated very successfully by surgery. The usual operation is to open the pulmonary stenosis, to close the septal defect, and to reconstruct the flow pathway into the aorta When successful, the average life expectancy increases from only a few years to 50 or more years

PULMONARY STENOSIS

Often a child is born with pulmonary stenosis but without other congenital abnormalities Pulmonary stenosis occasionally occurs which is so severe that the right side of the heart is likely to fail at an early age because blood flow from the right ventricle into the lungs is greatly impeded, the right ventricular systolic pressure rising to 75 to 100 mm Hg instead of the normal 26 mm Hg The right side of the heart dilates, and its muscle becomes greatly hypertrophied in order to withstand the load Also, a loud stenotic murmur is heard over the pulmonary valve area. In most cases of pulmonary stenosis, the stenotic area can be enlarged by surgical operation so that the heart resumes normal function

USE OF EXTRACORPOREAL CIRCULATION DURING CARDIAC SURGERY

It is almost impossible to repair intracardiac defects while the heart is still pumping Therefore, many different types of artificial *heart-lung machines* have been developed to take the place of the heart and lungs during the course of operation Such a system is called an *extracorporeal circulation* The system consists principally of (1) a pump and (2) an oxygenating device Almost any type of pump that does not cause hemolysis of the blood seems to be suitable

The different principles that have been used for oxygenating blood are (1) bubbling oxygen through the blood and then removing the bubbles from the blood before passing it back into the patient, (2) dripping the blood downward over the surfaces of large areas of plastic sheet in the presence of oxygen, (3) passing the blood over the surfaces of rotating discs, and (4) passing the blood between thin membranes or through thin tubes that are porous to oxygen and carbon dioxide

The different oxygenators have been fraught with many difficulties, including hemolysis of the blood, development of small clots in the blood, likelihood of small bubbles of oxygen or small emboli of antifoam agent passing into the arteries of the patient, necessity for large quantities of blood to prime the entire system, failure to exchange adequate quantities of oxygen, and the necessity to use heparin in the system to prevent blood coagulation, the heparin also preventing adequate hemostasis during the surgical procedure. Yet, despite difficulties, in the hands of experts patients can be kept on artificial heart-lung machines for many hours while operations are performed on the inside of the heart

HYPERTROPHY OF THE HEART IN VALVULAR AND CONGENITAL HEART DISEASE

Hypertrophy of cardiac muscle is one of the most important mechanisms by which the heart adapts to increased work loads, whether these loads be caused by increased pressure against which the heart muscle must contract or by increased volume that must be pumped Some investigators believe that it is the increased work load itself that causes the hypertrophy; others believe the increased metabolic rate of the muscle to be the primary stimulus and the work load simply to be the cause of the increase in the metabolic rate Regardless of which of these is correct, one can calculate approximately how much hypertrophy will occur in each chamber of the heart by multiplying ventricular output times the pressure against which the ventricle must work, with extra emphasis on the pressure. Thus, hypertrophy occurs in most types of valvular and congenital disease, as follows

In *aortic stenosis* and *aortic regurgitation*, the left ventricular musculature hypertrophies tremendously, sometimes to as much as four to five times normal, so that the weight of the heart on occasion may be as great as 800 grams instead of the normal 300 grams

In *mitral stenosis*, the left atrium hypertrophies and dilates, and the right ventricle hypertrophies slightly, but no left ventricular hypertrophy occurs. In *mitral regurgitation* moderate hypertrophy of the left ventricle occurs, and some hypertrophy of the right ventricle also develops owing to back pressure effects through the lungs causing elevation of the pulmonary arterial pressure

In *patent ductus arteriosus*, the work load of both ventricles is increased The left ventricle pumps an average of twice the quantity of blood that it normally pumps, therefore, it would be expected to hypertrophy. On the other hand, the right ventricle must pump its blood against a much higher than normal pulmonary arterial pressure because of the large quantity of blood refluxing from the aorta into the pulmonary artery Consequently, the right ventricle also hypertrophies

In *tetralogy of Fallot*, one of the cardinal signs is marked right ventricular hypertrophy, for the right ventricle must pump against the pressure in the aorta, and it also must pump an increased volume of blood On the other hand, the work load of the left ventricle is actually less than normal because of the reduced blood flow from the lungs into the left heart. Therefore, the right ventricle in tetralogy of Fallot is often larger than the normal left ventricle, while the left ventricle may be smaller than normal

Detrimental Effects of the Late Stages of Hypertrophy. Though physiological hypertrophy of heart muscle is usually very beneficial to cardiac function, extreme degrees of hypertrophy sometimes lead to failure. One of the reasons for this is that the coronary vasculature usually does not increase to the same extent as the mass of muscle. A second reason is that fibrosis often develops in the muscle, with fibrous tissue replacing degenerating muscle fibers. Because of the sometimes disproportionate increase in muscle mass relative to coronary flow, relative ischemia sometimes develops as the muscle hypertrophies, and coronary insufficiency easily ensues Therefore, anginal pain is a frequent accompaniment of many valvular and congenital heart diseases

REFERENCES

Agarwal, A. K.. Aortic stenosis Diagnosis and therapy of the elderly patient. *Geriatrics*, *40* 105, 110, 1985

Alpert, N (ed.) Cardiac Hypertrophy New York, Academic Press, 1971

Badeer, H. S · Development of cardiomegaly *Cardiology*, 57 247, 1972

Barry, A.. Aortic and Tricuspid Valvular Disease New York, Appleton-Century-Crofts, 1980

Bayer, L. M, and Honzik, M P Children with Intracardiac Defects. Springfield, Ill , Charles C Thomas, 1974

Bernstein, E F. (ed) Noninvasive Diagnostic Techniques in Vascular Disease St. Louis, C V Mosby, 1978.

Butterworth, J S , *et al* Cardiac Auscultation–Including Audio-Visual Principles, 2nd Ed New York, Grune & Stratton 1960

Comroe, J. H , Jr , and Dripps, R D The Top Ten Clinical Advances in Cardiovascular Pulmonary Medicine and Surgery Between 1945 and 1975, How They Came About Final Report Washington, U S Govt Printing Office, 1978

Cooley, D. A , and Hallman, G L Surgical Treatment of Congenital Heart Disease Philadelphia, Lea & Febiger, 1966

DePace, N L., *et al*.. Acute severe mitral regurgitation Pathophysiology, clinical recognition, and management *Am J Med* 76 293, 1985

Dieshrich, E B (ed) Noninvasive Cardiovascular Diagnosis Littleton MA, Wright, PSG, 1981

Elliott, L P.. An Angiocardiographic and Plain Film Approach to Complex Congenital Heart Disease Chicago Year Book Medical Publishers, 1978

Grossman, W. (ed) Cardiac Catheterization and Angiography, 2nd Ed Philadelphia, Lea & Febiger, 1980

Haft, J. I , and Horowitz, M S Clinical Echocardiography Mount Kisco, N Y , Futura Publishing Co , 1978

Hayase, S , and Murao, S Cardiology New York, Elsevier/North Holland, 1979

Heymann, M A , and Rudolph, A M Control of the ductus arteriosus *Physiol Rev*, *55* 62, 1975

Hugenholtz, P D , *et al* Current status of echocardiology *In* Dickinson, C J., and Marks, J (eds) Developments in Cardiovascular Medicine Lancaster, England, MTP Press 1978, p 15

Kidd, L., and Somerville, J Long-term survival after operations for congenital heart disease *In* Hayase, S , and Murao S (eds) Cardiology Proceedings of the VIII World Congress of Cardiology Tokyo, 1978 New York, Elsevier/North Holland, 1979, p 774

Kobayashi, T , *et al* (eds) Cardiac Adaptation Baltimore University Park Press, 1978

Kotler, M N , and Segal, B L (eds) Clinical Echocardiography Philadelphia, F A Davis Co , 1978

Kreel, L., Computed tomography and the cardiovascular system *In* Dickinson, C J , and Marks, J (eds) Developments in Cardiovascular Medicine Lancaster, England, MTP Press, 1978, p 3

Kremkau, F W Diagnostic Ultrasound, Physical Principles and Exercises New York, Grune & Stratton, 1980

Lear, M W : Heartsounds New York, Simon & Schuster, 1979

Lindsay, J , Jr , and Hurst, J W (eds) The Aorta New York, Grune & Stratton, 1979

Lundstrom, N (ed.). Echocardiography in Congenital Heart Disease New York, Elsevier/North-Holland, 1978

Mair, D D , and Ritter, D G.: The physiology of cyanotic congenital heart disease *Int Rev Physiol*., 9 275, 1976.

Mitchell, J H., and Wildenthal, K.. Static (isometric) exercise and the heart. Physiological and clinical considerations. *Annu Rev Med*, 25 369, 1974.

Nadas, A S (ed) Pulmonary Stenosis, Aortic Stenosis, Ventricular Septal Defect. Dallas, American Heart Association, 1977

Nanda, N C , and Gramiak, R.. Clinical Echocardiography St. Louis, C V Mosby, 1978

New York Heart Association Criteria Committee Nomenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels, 8th Ed Boston, Little, Brown, 1979

Nimura, Y., and Henry, W.. Newer aspects of echocardiography *In* Hayase, S , and Murao, S (eds.) Cardiology Proceedings of the VIII World Congress of Cardiology, Tokyo. 1978 New York, Elsevier/North Holland, 1979, p. 1053

Olley, P M , and Coceani, F Prostaglandins and the ductus arteriosus. *Annu Rev Med*, 32 375, 1981

Perloff, J K The Clinical Recognition of Congenital Heart Disease Philadelphia, W B Saunders Co., 1978

Pomerance, A , and Davis, M J (eds) Pathology of the Heart. Philadelphia, J B Lippincott Co , 1975

Rapaport, E (ed.) Current Controversies in Cardiovascular Disease *Philadelphia*, W B Saunders Co , 1980

Reiser S J The medical influence of the stethoscope *Sci Am*, *240*(2) 148, 1979

Roberts, W C , and Spray, T L Pericardial Heart Disease Chicago, Year Book Medical Publishers, 1977

Senning, A Surgical treatment of valvular disease *In* Hayase, S , and Murao, S (eds) Cardiology Proceedings of the VIII World Congress of Cardiology, Tokyo, 1978 New York, Elsevier/North-Holland, 1979, p 751

Shah, P M., and Roberts, D L Diagnosis and Treatment of Aortic Valve Stenosis Chicago, Year Book Medical Publishers, 1977

Stapleton J F , and Harvey, W P Heart sounds, murmurs, and precordial movements *In* Sodeman, W A , Jr , and Sodeman, T M (eds) Pathologic Physiology Mechanisms of Disease, 6th Ed Philadelphia, W B Saunders Co , 1979, p 335

Stollermann, G H , and Shokawa, Y Current status of rheumatic fever and rheumatic carditis *In* Hayase, S , and Murao, S (eds) Cardiology Proceedings of the VIII World Congress of Cardiology, Tokyo, 1978 New York, Elsevier/North Holland, 1979, p 500

Taussig, H Congenital Malformations of the Heart Vol 1 General Considerations, 2nd Ed Vol 2 Specific Malformations, 2nd Ed Cambridge, Mass., Harvard University Press, 1960

Weisfeldt, L., and Chandra, N Physiology of cardiopulmonary resuscitation. *Annu Rev Med*, *32* 435, 1981

Wynwood, R. S Essentials of Clinical Diagnosis in Cardiology Baltimore, E Arnold, 1981

Zak, R., and Rabinowitz, M Molecular aspects of cardiac hypertrophy *Annu Rev Physiol*, *41* 539, 1979

28

Circulatory Shock and Physiology of Its Treatment

Circulatory shock means generalized inadequacy of blood flow throughout the body, to the extent that the tissues are damaged because of too little flow, especially too little delivery of oxygen and other nutrients to the tissue cells. Even the cardiovascular system itself—the heart musculature, the walls of the blood vessels, the vasomotor system, and other circulatory parts—begins to deteriorate so that the shock becomes progressively worse.

THE PHYSIOLOGICAL CAUSES OF SHOCK

Circulatory Shock Caused by Decreased Cardiac Output

Shock usually results from inadequate cardiac output. Therefore, any factor that reduces the cardiac output will likely lead to circulatory shock. Basically, two types of factors can severely reduce the cardiac output, these are:

(1) *Cardiac abnormalities that decrease the ability of the heart to pump blood.* These include especially *myocardial infarction* but also *toxic states of the heart, severe heart valve dysfunction,* heart *arrhythmias,* and other conditions. The circulatory shock that results from diminished cardiac pumping ability is called *cardiogenic shock.* This was discussed in detail in Chapter 26, where it was pointed out that about 85 per cent of persons who develop cardiogenic shock do not survive.

(2) *Factors that decrease the venous return.* The most common cause of this is *diminished blood volume,* but venous return also can be reduced as a result of decreased *vasomotor tone* or *obstruction to blood flow* at some point in the circulation, especially in the venous return pathway to the heart.

Circulatory Shock Without Diminished Cardiac Output

Occasionally, the cardiac output is normal or even greater than normal and yet the person is in circulatory shock. This can result (a) from *excessive metabolism of the body so that even a normal cardiac output is inadequate,* or (b) from *abnormal tissue perfusion patterns so that most of the cardiac output is passing through blood vessels that are not supplying the local tissues with nutrition.* These conditions are seen most frequently in the type of shock called *septic shock,* which has often also been called "blood poisoning."

The different specific causes of shock will be discussed later in the chapter. For the time being, it is important to note that all of them lead to *inadequate delivery of nutrients to the tissues and inadequate removal of cellular waste products from the tissues.*

What Happens to the Arterial Pressure in Circulatory Shock?

In the minds of many physicians, the arterial pressure is the principal measure of the adequacy of circulatory function. However, the arterial pressure often can be seriously misleading, because many times a person may be in severe shock and still have a normal pressure because powerful nervous reflexes often keep the pressure from falling. Yet at other times the arterial pressure can fall to as low as one-half normal, but the person still has normal tissue perfusion and is not in shock.

Nevertheless, it is true that in most types of shock, especially that caused by severe blood loss, the arterial blood pressure usually does decrease at the same time that the cardiac output decreases, though usually not as much as the decrease in output.

The End-Stages of Circulatory Shock, Whatever the Cause

Once circulatory shock reaches a critical state of severity, regardless of its initiating cause, *the shock itself breeds more shock.* That is, the inadequate blood flow causes the circulatory system itself to begin to deteriorate. This in turn causes even more decrease in cardiac output, and a vicious circle ensues, with progressively increasing circulatory shock, still less adequate tissue perfusion, still more shock, and so forth until death. It is with this late stage of the circulatory shock that we are especially concerned, because appropriate physiologic treatment can often reverse the rapid slide to oblivion.

The Stages of Shock. Because the characteristics of circulatory shock change at different degrees of severity, shock is generally divided into three major stages:

(1) A *nonprogressive stage* (sometimes called the *com*

pensated stage) from which the normal circulatory compensatory mechanisms will eventually cause full recovery if the initiating cause does not become any worse.

(2) A *progressive stage*, in which the shock becomes steadily worse until death.

And (3) an *irreversible stage*, in which the shock has progressed to such an extent that all forms of known therapy will be inadequate to save the life of the person even though for the moment the person is still alive

Now, let us discuss the different stages of circulatory shock caused by decreased blood volume, which will illustrate the basic principles. Then we can consider the special characteristics of other initiating causes of shock

SHOCK CAUSED BY HYPOVOLEMIA—HEMORRHAGIC SHOCK →

Hypovolemia means diminished blood volume, and hemorrhage is perhaps the most common cause of hypovolemic shock.

Hemorrhage *decreases the mean systemic filling pressure* and as a consequence decreases venous return As a result, the cardiac output falls below normal, and shock ensues. Obviously, all degrees of shock can result from hemorrhage, from the mildest diminishment of cardiac output to almost complete cessation of output

RELATIONSHIP OF BLEEDING VOLUME TO CARDIAC OUTPUT AND ARTERIAL PRESSURE

Figure 28–1 illustrates the effect on both cardiac output and arterial pressure of removing blood from the circulatory system over a period of about half an hour. Approximately 10 per cent of the total blood volume can be removed with no significant effect on arterial pressure or cardiac output, but greater blood loss usually diminishes the cardiac output first and later the pressure, both of these falling to zero when about 35 to 45 per cent of the total blood volume has been removed

Sympathetic Reflex Compensation in Shock. Fortunately, the decrease in arterial pressure caused by blood loss initiates powerful sympathetic reflexes (initiated mainly by the baroreceptors) that stimulate the sympathetic vasoconstrictor system throughout the body, resulting in three important effects. (1) The arterioles constrict in most parts of the body, thereby greatly increasing the total peripheral resistance. (2) The veins and venous reservoirs constrict, thereby helping to maintain adequate venous return despite diminished blood volume. And (3) heart activity increases markedly, sometimes increasing the heart rate from the normal value of 72 beats per minute to as much as 200 beats per minute.

Value of the Reflexes. In the absence of the sympathetic reflexes, only 15 to 20 per cent of the blood volume can be removed over a period of half an hour before a person will die; this is in contrast to a 30 to 40 per cent loss of blood volume that a person can sustain when the reflexes are intact. Therefore, the reflexes extend the amount of blood loss that can occur without causing death to about two times that which is possible in their absence.

Greater Effect of the Reflexes in Maintaining

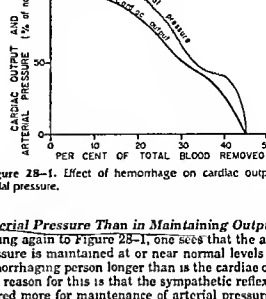


Figure 28–1. Effect of hemorrhage on cardiac output and arterial pressure.

Arterial Pressure Than in Maintaining Output. Referring again to Figure 28–1, one sees that the arterial pressure is maintained at or near normal levels in the hemorrhaging person longer than is the cardiac output. The reason for this is that the sympathetic reflexes are geared more for maintenance of arterial pressure than for maintenance of output They increase the arterial pressure mainly by increasing the total peripheral resistance, which has no beneficial effect on cardiac output; but the sympathetic constriction of the veins is important to keep venous return and cardiac output from falling too much.

Especially interesting is the second plateau in the arterial pressure curve of Figure 28–1. This results from activation of the central nervous system (CNS) ischemic response, which causes extreme stimulation of the sympathetic nervous system, as discussed in Chapter 21. This effect of the CNS ischemic response can be called the "last ditch stand" of the sympathetic reflexes in their attempt to keep the arterial pressure from falling too low.

Protection of Coronary and Cerebral Blood Flow by the Reflexes. A special value of the maintenance of normal arterial pressure even in the face of decreasing cardiac output is protection of blood flow through the coronary and cerebral circulatory systems. Sympathetic stimulation does not cause significant constriction of either the cerebral or cardiac vessels. In addition, in both these vascular beds local autoregulation is excellent, which prevents moderate decreases in arterial pressure from significantly affecting their blood flows. Therefore, blood flow through the heart and brain is maintained essentially at normal levels as long as the arterial pressure does not fall below about 70 mm Hg, despite the fact that blood flow in many other areas of the body might be decreased almost to zero by this time because of vasospasm.

NONPROGRESSIVE AND PROGRESSIVE HEMORRHAGIC SHOCK

Figure 28–2 illustrates an experiment that we performed in dogs to demonstrate the effects of different degrees of hemorrhage on the subsequent course of arterial pressure. The dogs were bled rapidly until their arterial pressures fell to different levels. Those dogs whose pressures fell immediately to no lower than 45 mm Hg (Groups I, II, and III) all eventually recovered; the recovery occurred rapidly if the

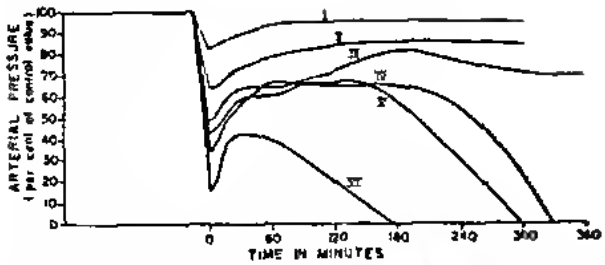


Figure 28–2. Course of arterial pressure in dogs after different degrees of acute hemorrhage. Each curve represents the average results from six dogs.

slightly (Group I) but occurred slowly if it fell almost to the 45 mm Hg level (Group III). On the other hand, when the arterial pressure fell below 45 mm Hg (Groups IV, V, and VI), all the dogs died, though many of them hovered between life and death for many hours before the circulatory system began to deteriorate.

This experiment demonstrates that the circulatory system can recover as long as the degree of hemorrhage is no greater than a certain critical amount. However, crossing this critical amount by even a few milliliters of blood loss makes the eventual difference between life and death. Thus, hemorrhage beyond a certain critical level causes shock to become *progressive*. That is, *the shock itself causes still more shock*, the condition becoming a vicious circle that leads eventually to complete deterioration of the circulation and to death.

Nonprogressive Shock—Compensated Shock

If shock is not severe enough to cause its own progression, the person eventually recovers. Therefore, shock of this lesser degree can be called *nonprogressive shock*. It is also frequently called *compensated shock*, meaning that the sympathetic reflexes and other factors have compensated enough to prevent deterioration of the circulation.

The factors that cause a person to recover from moderate degrees of shock are the negative feedback control mechanisms of the circulation that attempt to return cardiac output and arterial pressure to normal levels. These include:

1. The *baroreceptor reflexes*, which elicit powerful sympathetic stimulation of the circulation.
2. The *central nervous system ischemic response*, which elicits even more powerful sympathetic stimulation throughout the body but is not activated until the arterial pressure falls below 60 mm Hg.
3. *Reverse stress-relaxation of the circulatory system*, which causes the blood vessels to contract down around the diminished blood volume so that the blood volume that is available will more adequately fill the circulation.
4. *Formation of angiotensin*, which constricts the peripheral arteries and causes increased conservation of water and salt by the kidneys, both of which help prevent progression of the shock.
5. *Formation of vasopressin (antidiuretic hormone)*, which constricts the peripheral arteries and veins and also greatly increases water retention by the kidneys.
6. *Compensation mechanisms that return the blood volume back toward normal*, including absorption of large quantities of fluid from the intestinal tract, absorption of fluid from the interstitial spaces of the body, conservation of water and salt by the kidneys, and increased thirst and increased appetite for salt which make the person drink water and eat salty foods if abl

The sympathetic reflexes provide immediate help t ward bringing about recovery, for they become max mally activated within 30 seconds after hemorrhag The angiotensin and vasopressin mechanisms, as we as the reverse stress-relaxation that causes contractio of the blood vessels and venous reservoirs around th blood, all require 10 minutes to an hour to occur com pletely, but, nevertheless, these aid greatly in increasir the arterial pressure or increasing the mean system filling pressure and thereby increasing the return blood to the heart. Finally, the readjustment of bloc volume by absorption of fluid from the interstitial space and from the intestinal tract, as well as the ingestio and absorption of additional quantities of fluid and sal may require from 1 to 48 hours, but eventually recover takes place provided the shock does not become sever enough to enter the progressive stage.

Progressive Shock - The Vicious Circle of Cardiovascular Deterioration

Once shock has become severe enough, the structur of the circulatory system themselves begin to deteri rate, and various types of positive feedback develop tha can cause a vicious circle of progressively decreasin cardiac output. Figure 28–3 illustrates some of the different types of positive feedback that further depre the cardiac output in shock. These are the following

Cardiac Depression. When the arterial pressu falls low enough, *coronary blood flow decreases belo that required for adequate nutrition of the myocardiu* itself. This obviously weakens the heart and theret decreases the cardiac output still more. As a conse quence, the arterial pressure falls still further, and th coronary blood flow decreases more, making the hea still weaker. Thus, a positive feedback cycle has deve oped whereby the shock becomes more and more sever

Figure 28–4 illustrates cardiac output curves fro experiments in dogs, showing progressive deteriorati of the heart at different times following the onset shock. A dog was bled until the arterial pressure fell 30 mm Hg, and the pressure was held at this level further bleeding or retransfusion of blood as require Note that there was little deterioration of the hea during the first two hours, but by four hours the hea had deteriorated about 40 per cent; then, rapidly, durir the last hour of the experiment the heart deteriorate almost completely.

Thus, one of the important features of progressi shock, whether it be hemorrhagic in origin or caused any other way, is eventual progressive deterioration the heart. In the early stages of shock, this plays ve little role in the condition of the person, partly becaus deterioration of the heart itself is not very severe durin the first hour or so of shock but mainly because th heart has tremendous reserve that makes it normall capable of pumping 300 to 400 per cent more blood tha is required by the body for adequate nutrition. Howeve in the very latest stages of shock, deterioration of th heart is probably the most important factor in th further progression of the shock.

Vasomotor Failure. In the early stages of shoc various circulatory reflexes cause intense activity of th

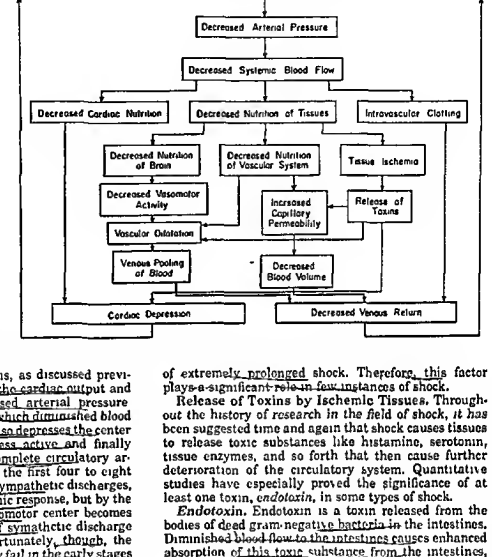


Figure 28–3. Different types of feedback that can lead to progression of shock.

sympathetic nervous system. This, as discussed previously, helps delay depression of the cardiac output and especially helps prevent decreased arterial pressure. However, there comes a point at which diminished blood flow to the vasomotor center itself so depresses the center that it becomes progressively less active and finally totally inactive. For instance, complete circulatory arrest to the brain causes, during the first four to eight minutes, the most intense of all sympathetic discharges, resulting in a typical CNS ischemic response, but by the end of 10 to 15 minutes the vasomotor center becomes so depressed that no evidence of sympathetic discharge at all can be demonstrated. Fortunately, though, the vasomotor center does not usually fail in the early stages of shock—only in the late stages.

Thrombosis of the Minute Vessels—Sludged Blood. Recently, several research workers have shown that thrombosis occurring in many of the minute vessels in the circulatory system can be one of the causes of shock progression. That is, blood flow through many tissues becomes extremely sluggish, but tissue metabolism continues so that large amounts of acid, either carbonic acid or lactic acid, continue to empty into the local blood vessels and to increase greatly the acidity of the blood. This acid, plus other deterioration products from the ischemic tissues, causes blood agglutination or actual blood clots, thus leading to minute plugs in the small vessels. Even if the vessels do not become plugged, the tendency for the cells to stick to each other makes it more difficult for blood to flow through the microvasculature, giving rise to the term *sludged blood*.

Increased Capillary Permeability. After many hours of capillary hypoxia the permeability of the capillaries gradually increases and large quantities of fluid begin to transude into the tissues. This further decreases the blood volume, with resultant further decrease in cardiac output, thus making the shock still more severe.

Fortunately, capillary hypoxia does not cause increased capillary permeability until the very late stages of extremely prolonged shock. Therefore, this factor plays a significant role in few instances of shock.

Release of Toxins by Ischemic Tissues. Throughout the history of research in the field of shock, it has been suggested time and again that shock causes tissues to release toxic substances like histamine, serotonin, tissue enzymes, and so forth that then cause further deterioration of the circulatory system. Quantitative studies have especially proved the significance of at least one toxin, *endotoxin*, in some types of shock.

Endotoxin. Endotoxin is a toxin released from the bodies of dead gram-negative bacteria in the intestines. Diminished blood flow to the intestines causes enhanced absorption of this toxic substance from the intestines, and it then causes extensive vascular dilatation, greatly

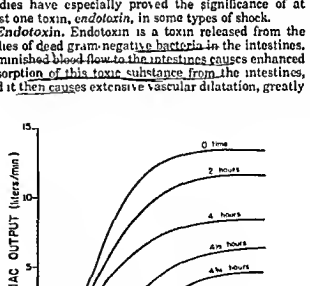


Figure 28–4. Function curves of the heart at different times after hemorrhagic shock begins. (These curves are extrapolated to the human heart from data obtained in dog experiments by Dr J W Crowell.)

increased cellular metabolism despite the inadequate nutrition of the cells, and cardiac depression. Though this toxin can play a major role in some types of shock, especially septic shock discussed later in the chapter, it still is not clear how much endotoxin is released during hemorrhagic shock and whether or not it is an especially important factor in the progression of this type of shock.

Generalized Cellular Deterioration. As shock becomes very severe, many signs of generalized cellular deterioration occur throughout the body. One organ especially affected is the *liver*, primarily because of the lack of enough nutrients to support the normally high rate of metabolism in liver cells, but also partly because of the extreme vascular exposure of the liver cells to any toxic or other abnormal metabolic factors in shock. Among the different effects that are known to occur are:

1. Active transport of sodium and potassium through the cell membrane is greatly diminished. As a result, sodium and chloride accumulate in the cells and potassium is lost from the cells. In addition the cells begin to swell.
2. Mitochondrial activity in the liver cells, as well as in many other tissues of the body, becomes severely depressed.
3. Lysosomes begin to split in widespread tissue areas, with intracellular release of hydrolases that cause further intracellular deterioration.
4. Cellular metabolism of nutrients such as glucose eventually becomes greatly depressed in the last stages of shock. The activities of some hormones are depressed as well, including as much as 200-fold depression in the action of insulin.

Obviously, all these effects contribute to further deterioration of many different organs of the body, including especially (1) the *liver*, with depression of its many metabolic and detoxification functions, (2) the lungs, with eventual development of pulmonary edema and poor ability to oxygenate the blood, and (3) the heart, thereby further depressing the contractility of the heart.

Tissue Necrosis in Severe Shock—Effect of Pattern of Vascular Blood Flow in Different Organs. Not all cells of the body are equally damaged by shock because some tissues have better blood supplies than others. For instance, the cells adjacent to the arterial ends of capillaries receive better nutrition than the cells adjacent to the venous ends of the same capillaries. Therefore, one would expect more nutritive deficiency around the venous ends of capillaries than elsewhere. This is precisely the effect that Crowell has found in studying tissue areas in many parts of the body. For instance, Figure 28–5 illustrates necrosis in the center of a liver lobule, the portion of the lobule that is last to be bathed by the blood as it passes through the liver sinusoids.

Similar punctate lesions occur in heart muscle, though here a definite repetitive pattern such as occurs in the liver cannot be demonstrated. Nevertheless, the cardiac lesions probably play an important role in leading to the final irreversible stage of shock. Deteriorative lesions also occur in the kidneys, especially in the kidney tubules, leading to kidney failure and subsequent uremic death several days later. And deterioration of lungs often leads to respiratory distress and death days later—called the *shock lung syndrome*.

Acidosis in Shock. Most of the metabolic derangements that occur in shocked tissue can lead to acidosis. Especially important is the poor delivery of oxygen to

Figure 28–5. Necrosis of the central portion of a liver lobule severe shock. (Courtesy of Dr. Jack Crowell.)

the tissues, which greatly diminishes oxidative metabolism of the foodstuffs. When this occurs, the cells obtain their energy by the anaerobic process of glycolysis that leads to tremendous quantities of excess *lactic acid* in the blood. In addition, the poor blood flow through the tissues prevents normal removal of carbon dioxide, the carbon dioxide reacts locally in the cells with water to form very high concentrations of intracellular carbonic acid, this in turn reacts with the various tissue buffers to form still other intracellular acidic substances.

Thus, another deteriorative effect of shock is both generalized and local tissue acidosis, leading to still further progression of the shock itself.

Positive Feedback and the Vicious Circle of Progressive Shock

All the factors just discussed that can lead to further progression of shock are types of *positive feedback*. That is, each increase in the degree of shock causes a further increase in the shock.

However, positive feedback does not always necessarily lead to a vicious circle. Whether or not a vicious circle develops depends on the intensity of the positive feedback. In mild degrees of shock, the negative feedback mechanisms—the sympathetic reflexes, the stress-relaxation mechanism of the blood reservoirs, the absorption of fluid into the blood from the interstitial spaces, and others—can easily overcome the positive feedback influences and can therefore cause recovery. In severe degrees of shock, however, the positive feedback mechanisms become more and more powerful, thus leading to such rapid deterioration of the circulation that the negative feedback systems cannot return the cardiac output to normal.

Thus, considering once again the principles of positive feedback and vicious circles as discussed in Chapter 1, one can readily understand why there is a critical cardiac output level above which a person in shock recovers and below which the person enters a vicious circle of circulatory deterioration.

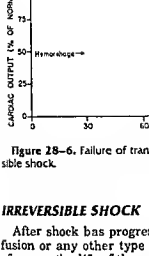


Figure 28–6. Failure of transfusion to prevent death in irreversible shock.

IRREVERSIBLE SHOCK

After shock has progressed to a certain stage, transfusion or any other type of therapy becomes incapable of saving the life of the person. Therefore, the person is then said to be in the *irreversible stage of shock*. Ironically, even in this irreversible stage, therapy can on occasion still return the arterial pressure and even the cardiac output to normal for short periods of time, but the circulatory system nevertheless continues to deteriorate and death ensues in another few minutes to few hours.

Figure 28–6 illustrates this effect, showing that transfusion during the irreversible stage can sometimes cause the cardiac output (as well as the arterial pressure) to return to normal. However, the cardiac output soon begins to fall again, and subsequent transfusions have less and less effect. Thus, something has changed in the overall function of the circulatory system that may not necessarily affect the *immediate* ability of the heart to pump blood but over a long period of time does depress this ability and results in death. Now the question remains: What factor or factors lead to the eventual total deterioration of circulatory function?

The answer to this question seems to be, simply, that beyond a certain point so much tissue damage has occurred, so many destructive enzymes have been released into the body fluids, so much acidosis has developed, and so many other destructive factors are now in progress that even a normal cardiac output cannot reverse the continuing deterioration. Therefore, in severe shock a stage is eventually reached beyond which the person is destined to die even though vigorous therapy can still return the cardiac output to normal for short periods of time.

Depletion of Cellular High Energy Phosphate Reserves in Irreversible Shock. The high energy phosphate reserves in the tissues of the body, especially in the liver and in the heart, are greatly diminished in severe degrees of shock. Essentially all the creatinine phosphate is degraded, and almost all the ATP has been degraded to ADP, AMP, or adenosine. Much of the adenosine that is derived from degradation of the ATP diffuses out of the cells into the circulating blood and is converted into uric acid, a substance that cannot reenter the cells to reconstitute the adenosine phosphate system. Unfortunately, new adenosine can be synthesized at the rate of only about 2 per cent an hour, meaning that, once depleted, the high energy phosphate stores of the cells are difficult to replenish. Therefore, one of the most important end-results of deterioration in shock, and one that is perhaps the most significant of all in the development of the final state of irreversibility, is this cellular depletion of the high energy compounds.

Deterioration of the Heart as a Principal Cause of Irreversibility. It is clear that deterioration can occur in many different organ systems in shock and that the degeneration in any of these systems could become so severe that it would eventually be incompatible with continued life. However, in most instances it is probably deterioration of the heart itself that makes the shock irreversible. The reason for believing this is the following: Modern therapy is very effective in producing adequate venous return. Administration of blood and other substitution fluids can almost always provide adequate inflow pressure to the heart even in the most severe degrees of shock. But, still, in the late stages of shock, the heart fails to pump this inflowing blood. Therefore, the heart is the final weak link in the system.

On the other hand, if one does not use all forms of available therapy, a person can die of shock because of peripheral abnormalities such as continued loss of fluid into the tissues, pooling of blood in greatly distended blood vessels, respiratory failure, acidosis, and so forth. Unfortunately, in such instances, the patient dies of shock that is still reversible if adequate therapy is provided.

Special Role of Oxygen Deficiency in Shock and in Irreversibility

Though poor blood flow leads to tissue deterioration because of deficiency of many different nutrients, deficiency of oxygen almost certainly is the most important. For instance, in a large series of dogs, Crowell measured the accumulated deficit of oxygen usage of animals (a) in mild shock, (b) in moderate shock, and (c) in very severe shock. In some animals, this deficit accumulated slowly, whereas in others it accumulated very rapidly. But in each group of animals, when the average accumulated oxygen deficit reached 120 milliliters of oxygen per kilogram of body mass, 50 per cent of the animals died regardless of how long it took to accumulate this amount of oxygen deficit.

Therefore, it seems to be clear that the one most important nutrient necessary to prevent cellular deterioration and death during shock is oxygen.

HYPOVOLEMIC SHOCK CAUSED BY PLASMA LOSS

Loss of plasma from the circulatory system, even without the loss of whole blood, can sometimes be severe enough to reduce the total blood volume markedly, in this way causing typical hypovolemic shock similar in almost all details to that caused by hemorrhage. Severe plasma loss occurs in the following conditions:

Intestinal obstruction is often a cause of reduced plasma volume. The resulting distention of the intestine causes fluid to leak from the intestinal capillaries into the intestinal walls and intestinal lumen. This loss of fluid might result from elevated capillary pressure caused by increased resistance in the stretched veins over the surface of the intestine, or it might be caused by direct damage to the capillaries themselves. Nevertheless, the lost fluid has a very high content of protein, thereby reducing the total plasma protein as well as the plasma volume.

2 Often, in patients who have *severe burns* or other denuding conditions of the skin, so much plasma is lost through the exposed areas that the plasma volume becomes markedly reduced.

The hypovolemic shock that results from plasma loss has almost the same characteristics as the shock caused by hemorrhage, except for one additional complicating factor—the blood viscosity increases greatly as a result of plasma loss, and this further exacerbates the sluggishness of blood flow.

3 Loss of fluid from all fluid compartments of the body is called *dehydration;* this, too, can reduce the blood volume and cause hypovolemic shock very similar to that resulting from hemorrhage Some of the causes of this type of shock are (a) excessive sweating, (b) fluid loss in severe diarrhea or vomiting; (c) excess loss of fluid by nephrotic kidneys, (d) inadequate intake of fluid and electrolytes, and (e) destruction of the adrenal cortices, with consequent failure of the kidneys to reabsorb sodium, chloride, and water because of the absence of the hormone aldosterone

HYPOVOLEMIC SHOCK CAUSED BY TRAUMA

One of the most common causes of circulatory shock is trauma to the body Often the shock results simply from hemorrhage caused by the trauma, but it can also occur even without hemorrhage, for contusion of the body can often damage the capillaries sufficiently to allow excessive loss of plasma into the tissues This results in greatly reduced plasma volume with resultant hypovolemic shock Thus, whether or not hemorrhage occurs when a person is severely traumatized, the blood volume can still be markedly reduced.

The *pain* associated with serious trauma can be an additional aggravating factor in traumatic shock, for pain sometimes strongly inhibits the vasomotor center, thereby increasing the vascular capacitance and reducing the venous return. Various attempts have also been made to implicate toxic factors released by the traumatized tissues as one of the causes of shock following trauma However, cross-transfusion experiments have failed to show any such toxic element.

In summary, traumatic shock seems to result mainly from hypovolemia, though there may also be a moderate degree of concomitant neurogenic shock caused by the pain

NEUROGENIC SHOCK—INCREASED VASCULAR CAPACITY

Occasionally shock results without any loss of blood volume whatsoever Instead, the *vascular capacity* increases so much that even the normal amount of blood becomes incapable of adequately filling the circulatory system. One of the major causes of this is *loss of vasomotor tone* throughout the body especially causing massive dilation of the veins, and the resulting condition is then known as *neurogenic shock.*

The relationship of vascular capacity to blood volume . discussed in Chapter 18, where it was pointed out at either an increase in vascular capacity or a decrease in blood volume *reduces the mean systemic filling pressure*, which in turn reduces the venous return to the heart Diminished venous return caused by vasc dilatation is often called "venous pooling" of blood

Causes of Neurogenic Shock. Some of the different factors that can cause loss of vasomotor tone include

1 *Deep general anesthesia* often depresses the vasomotor center enough to cause vasomotor collapse, with resulting neurogenic shock.

2 *Spinal anesthesia*, especially when this extends all the way up the spinal cord, blocks the sympathetic outflow from the nervous system and is a common cause of neurogenic shock.

3 *Brain damage* is often a cause of vasomotor collapse. Many patients who have had brain concussion or contusion of the basal regions of the brain develop profound neurogenic shock. Also, even though short periods of medullary ischemia cause extreme vasomotor activity, prolonged ischemia can cause inactivation of the vasomotor neurons and can cause development of severe neurogenic shock

4 *Depressed vasomotor center* frequently occurs during fever, after excessive loss of sleep, or in metabolic disturbances This often leads to *fainting*, in which the peripheral blood vessels become greatly dilated As a result, blood pools, and the cardiac output falls drastically Such a person held in an upright position will go into the progressive stage of shock and can die as a result Fortunately, on fainting, a person usually falls to a horizontal position so that almost normal cardiac output ordinarily re-ensues almost immediately.

Vasovagal Syncope—Emotional Fainting. The circulatory collapse that results from "emotional" fainting usually is not caused by vasomotor failure but instead by strong emotional excitation of the parasympathetic nerves to the heart and of the vasodilator nerves to the skeletal muscles, thereby slowing the heart and reducing the arterial pressure. Therefore, the fainting that results from an emotional disturbance is called *vasovagal syncope* to differentiate it from the other types of fainting which result from diminished sympathetic activity throughout the body or from other causes of reduced cardiac output

ANAPHYLACTIC SHOCK

"Anaphylaxis" is an allergic condition in which the cardiac output and arterial pressure often fall drastically. This was discussed in Chapter 6. It results primarily from an antigen-antibody reaction that takes place all through the body immediately after an antigen to which the person is sensitive has entered the circulatory system. Such a reaction is detrimental to the circulatory system in several important ways. First, if the antigen-antibody reaction takes place in direct contact with the vascular walls or cardiac musculature, damage to these tissues presumably can result directly. Second, cells damaged anywhere in the body by the antigen-antibody reaction release several highly toxic substances into the blood Among these is *histamine* or *histamine-like substance*, released mainly from the circulating basophils and the mast cells outside the capillaries The histamine in turn causes (1) an increase in vascular capacity because of venous dilatation, (2) dilatation of the arterioles with resultant greatly reduced arterial pressure, and (3) greatly increased capillary permeability with rapid loss of fluid into the tissue spaces. Unfortunately, all the precise relationships of

the above factors in anaphylaxis have not been determined, but the sum total is a great reduction in venous return and often such serious shock that the person dies within minutes.

Intravenous injection of large amounts of histamine causes "histamine shock," which has characteristics almost identical with those of anaphylactic shock, though usually less severe

SEPTIC SHOCK

The condition that was formerly known by the popular name of "blood poisoning" is now called *septic shock* by most clinicians. This simply means widely disseminated infection in many areas of the body, with the infection being borne through the blood from one tissue to another and causing extensive damage Actually, there are many different varieties of septic shock because of the many different types of bacterial infection that can cause it and also because infection in one part of the body will produce different effects from those caused by infection elsewhere in the body.

Septic shock is extremely important to the clinician because it is this type of shock that, more frequently than any other kind of shock besides cardiogenic shock, causes patient death in the modern hospital Some of the typical causes of septic shock include

1. Peritonitis caused by spread of infection from the uterus and fallopian tubes, frequently resulting from instrumental abortion

2. Peritonitis resulting from rupture of the gut, sometimes caused by intestinal disease and sometimes by wounds.

3. Generalized infection resulting from spread of a simple skin infection such as streptococcal or staphylococcal infection.

4. Generalized gangrenous infection resulting specifically from gas gangrene bacilli, spreading first through the tissues themselves and finally by way of the blood to the internal organs, especially to the liver

5. Infection spreading into the blood from the kidney or urinary tract, often caused by colon bacilli

Special Features of Septic Shock. Because of the multiple types of septic shock, it is difficult to categorize this condition. However, some features often seen in septic shock are the following:

1. High fever.

2. Marked vasodilatation throughout the body, especially in the infected tissues

3. High cardiac output in perhaps half of the patients, caused by vasodilatation in the infected tissues and also by high metabolic rate and vasodilatation elsewhere in the body resulting from bacterial toxin stimulation of cellular metabolism and from the high body temperature.

4. Sludging of the blood, presumably caused by red cell agglutination in response to degenerating tissues

5. Development of microclots in widespread areas of the body, a condition called *disseminated intravascular coagulation*. Also, this causes the clotting factors to be used up so that hemorrhages occur into many tissues, especially into the gut wall and into the intestinal tract.

In the early stages of septic shock the patient usually does not have *signs of circulatory collapse* but, instead, only signs of the bacterial infection itself. However, as the infection becomes more severe, the circulatory system usually becomes involved either directly or as a secondary result of toxins from the bacteria, and *there finally comes a point at which deterioration of the circulation becomes progressive in the same way that progression occurs in all other types of shock. Therefore, the end stages of septic shock are not greatly different from the end stages of hemorrhagic shock,* even though the initiating factors are markedly different in the two conditions.

Endotoxin Shock. A special type of septic shock is known as endotoxin shock. It frequently occurs when a large segment of the gut becomes strangulated and loses most of its blood supply. The gut rapidly becomes gangrenous, and the bacteria in the gut multiply rapidly. Most of these bacteria are so-called "gram-negative" bacteria, mainly colon bacilli, that contain a toxin called *endotoxin* Another condition that also frequently causes colon bacilli septicemia is extension of urinary tract infections into the blood.

On entering the circulation endotoxin causes an effect very similar to that of anaphylaxis, sometimes resulting in severe shock Indeed, the cause of this shock is possibly an anaphylactic reaction caused by the endotoxin itself. However, endotoxin has direct effects on the circulation as well, including (1) toxic depression of the heart and (2) vascular dilatation

Severe shock also occurs during the course of diseases caused by other gram-negative bacterial infections, such as dysentery, tularemia, brucellosis, and typhoid fever. It is likely that this shock results at least partially from endotoxin released from the gram negative bacteria causing these conditions

EFFECTS OF SHOCK ON THE BODY

Decreased Metabolism in Hypovolemic Shock. The decreased cardiac output in hypovolemic shock reduces the amount of oxygen and other nutrients available to the different tissues, and this in turn reduces the level of metabolism that can be maintained by the different cells of the body. Usually a person can continue to live for only a few hours if the cardiac output falls to as low as 40 per cent of normal.

Muscular Weakness. One of the earliest symptoms of shock is severe muscular weakness which is also associated with profound and rapid fatigue whenever patients attempt to use their muscles. This obviously results from the diminished supply of nutrients—especially oxygen—to the muscles.

Body Temperature. Because of the depressed metabolism in shock, the amount of heat liberated in the body is reduced (except in septic shock in which the infection may cause an opposite effect). As a result, the body temperature tends to decrease if the body is exposed to even the slightest cold.

Mental Function. In the early stages of shock the person is usually conscious, though signs of mental haziness may be noted. As the shock progresses, the person falls into a state of stupor, and in the last stages of shock even the subconscious mental functions, including vasomotor control and respiration, fail.

A person who recovers from shock usually exhibits no permanent impairment of mental functions. However, following complete circulatory arrest, in which the blood flow is completely cut off from the brain for more than five minutes, the brain does often suffer severe perma-

nent impairment, the cause of which is discussed in a separate section on circulatory arrest later in the chapter.

Reduced Renal Function. The very low blood flow during shock greatly diminishes urine output or even causes cessation of output, because glomerular pressure falls below the critical value required for filtration of fluid into Bowman's capsule, as explained in Chapters 34 and 35 Also, the kidney has such a high rate of metabolism and requires such large amounts of nutrients that the reduced blood flow often causes *tubular necrosis*, which means death of the tubular epithelial cells, with subsequent sloughing and blockage of the tubules, causing total loss of function of the respective nephrons This is often a serious aftereffect of shock that has occurred during a major surgical operation; the patient sometimes survives the shock associated with the surgical procedure and then dies a week or so later of uremia

PHYSIOLOGY OF TREATMENT IN SHOCK

REPLACEMENT THERAPY

Blood and Plasma Transfusion. If a person is in shock caused by hemorrhage, the best possible therapy is usually transfusion of whole blood. If the shock is caused by plasma loss, the best therapy is administration of plasma; when dehydration is the cause, administration of the appropriate electrolytic solution can correct the shock

Unfortunately, whole blood is not always available, such as under battlefield randitions However, plasma can usually substitute adequately for whole blood because it increases the blood volume and restores normal hemodynamics Plasma cannot restore a normal hematocrit, but the human being can usually stand a decrease in hematocrit to about one-third normal before serious ransequences result if the cardiac output is adequate Therefore, in acute conditions it is reasonable to use plasma in place of whole blood for treatment of hemorrhagic and most other types of hypovolemic shock

Sometimes plasma also is unavailable In these instances, various *plasma substitutes* have been developed that perform almost exactly the same hemodynamic functions as plasma One of these is the following:

Dextran Solution as a Plasma Substitute. The principal requirement of a truly effective plasma substitute is that it remain in the circulatory system—that is, not filter through the capillary pores into the tissue spaces But, in addition, the solution must be nontoxic and must rantain appropriate electrolytes to prevent derangement of the extracellular fluid electrolytes on administration To remain in the circulation the plasma substitute must rantain some substance that has a large enough molecular size to exert colloid osmotic pressure

One of the most satisfactory substances that has been developed thus far for this purpose is dextran, a large polysaccharide polymer of glucose Certain bacteria se- te dextran as a by-product of their growth, and dextran is manufactured by a bacterial lture procedure. By varying the growth conditions of e bacteria, the molecular weight of the dextran can be rantrolled to the desired value. Dextrans of appropriate molecular size do not pass through the capillary pores and, therefore, can replace plasma proteins colloid osmotic agents.

Fortunately, few toxic reactions have been obser when using dextran to provide colloid osmotic pressu therefore, solutions of this substance have proved t a satisfactory substitute for plasma in much fluid placement therapy

TREATMENT OF SHOCK WITH SYMPATHOMIMETIC AND SYMPATHOLYTIC DRUGS

A *sympathomimetic drug* mimics sympathetic stim lation These drugs include norepinephrine, epinep rine, and a large number of long-acting drugs that ha the same effects as epinephrine and norepinephrine *sympatholytic drug*, on the other hand, blocks the acti of normal sympathetic stimuli on the circulatory syste

Both sympathomimetic and sympatholytic drugs ha been used in treating different types of shock, includi hemorrhagic shock However, there is no proof th either plays a very beneficial role in this type of sho The sympathomimetic drugs usually have little bene because the circulatory reflexes are usually alre maximally activated in severe hypovolemic shock a are already causing marked endogenous secretion norepinephrine and epinephrine On the other hand, t proponents of the use of sympatholytic drugs ha claimed that these drugs increase blood flow throu the tissues and thereby prevent nutritional dama Unfortunately, these drugs also usually cause paraly of the vasculature, resulting in increased pooling blood in the venous system, which reduces the card output and negates much if not all of the direct effect the drugs in increasing local blood flow. Yet a comb nation of *transfusion plus sympatholytic drugs* probab is beneficial in the treatment of many patients wi shock; the transfusion fills the expanded vascular b and prevents the pooling, while the sympatholytic dr allows rapid blood flow to the tissues.

OTHER THERAPY

Treatment by the Head-Down Position. When t pressure falls too low in most types of shock, especial hemorrhagic and neurogenic shock, placing the patie with the head as much as a foot lower than the feet w help tremendously in promoting venous return a thereby increasing cardiac output Therefore, this is t first essential in the treatment of many types of shoc

Oxygen Therapy. Since the major deleterious effe of most types of shock is too little delivery of oxygen the tissues, giving the patient oxygen to breathe can of benefit in some instances. However, this frequent is of far less value than one might expect because t problem usually is not inadequate oxygenation of t blood in the lungs, but inadequate transport of the blo after it is oxygenated

Treatment with Glucocorticoids. Glucocorticoi are frequently given to patients in severe shock f several reasons: (1) experiments have shown empirical that glucocorticoids frequently increase the strength the heart in the late stages of shock; (2) glucocorticoi stabilize the lysosomal membranes and prevent relea of lysosomal enzymes into the cytoplasm of the cell thus preventing deterioration from this source; (3) gl

corticoids might also aid in the metabolism of glucose by the severely damaged cells.

Treatment of Neurogenic Shock. In neurogenic shock, sympathomimetic drugs actually reverse the basic cause of the shock itself by increasing the vasomotor tone throughout the body, but even in this condition these drugs may not be required, for, as long as the patient is in the *horizontal or head-down position*, adequate venous return to the heart usually occurs and maintains a *sufficient cardiac output to prevent progressive shock.*

Treatment of Anaphylactic Shock. Unfortunately, anaphylactic shock occurs so rapidly that one often cannot institute any therapy before death ensues, but if therapy can be instituted the condition can often be ameliorated or almost completely reversed by rapid administration of epinephrine or some other sympathomimetic drug. This does not correct the basic cause of the anaphylaxis, but it does cause vasoconstriction, which opposes the vasodilatation caused by histamine in anaphylaxis. Also, if one suspects that anaphylaxis might occur in a patient, its severity can usually be reduced by preliminary administration of glucocorticoids, which attenuate the allergic reaction responsible for the anaphylaxis, or *preliminary administration of antihistaminics*, which reduce the effects of the histamine released during anaphylaxis. However, in both these instances the therapy must be given *before* anaphylaxis takes place, for which reason they are usually of little value in therapy of anaphylaxis

Pain-Relieving Drugs in Traumatic Shock. The use of pain-relieving drugs, such as morphine, has been found clinically to reduce the severity of most traumatic shock, supposedly by removing the neurogenic element of the shock.

CIRCULATORY ARREST

A condition closely allied to circulatory shock is *circulatory arrest, in which all blood flow completely stops.* This occurs frequently on the surgical operating table as a result of *cardiac arrest* or of *ventricular fibrillation.*

Ventricular fibrillation can usually be stopped by strong electroshock of the heart, the basic principles of which were described in Chapter 14.

Cardiac arrest usually results from too little oxygen in the anesthetic gaseous mixture or from a depressant effect of the anesthesia itself A normal cardiac rhythm can usually be restored by removing the anesthetic and then applying cardiopulmonary resuscitation procedures for a few minutes while supplying the patient's lungs with adequate quantities of ventilatory oxygen

EFFECT OF CIRCULATORY ARREST ON THE BRAIN

A special problem in circulatory arrest is to prevent detrimental effects in the brain as a result of the arrest. In general, four to five minutes of circulatory arrest causes permanent brain damage in over half the patients, and circulatory arrest for as long as 10 minutes almost universally destroys most, if not all, of the mental powers.

For many years it has been taught that these detrimental effects on the brain were caused by the cerebral hypoxia that occurred during circulatory arrest. However, recent studies have shown that dogs can almost universally stand up to 30 minutes of circulatory arrest without permanent brain damage *if the blood is removed from the brain circulation prior to the arrest.* On the basis of these and other studies, it is postulated that the circulatory arrest causes vascular *clots* to develop throughout the brain and that these cause permanent or semipermanent ischemia of brain areas. This accords well with the results in human beings who have undergone long periods of circulatory arrest, for complete destruction of large areas in one side of the brain often *occurs while corresponding areas in the opposite side of* the brain, which should also be affected if hypoxia were the cause of the damage, are not affected even in the slightest.

REFERENCES

Bernton E W, *et al* Opioids and neuropeptides Mechanisms in circulatory shock *Fed. Proc , 44* 290, 1985

Bond, R. F, and Johnson, G , III Vascular adrenergic interactions during hemorrhagic shock *Fed Proc , 44* 281, 1985

Crowell J W Cardiac deterioration as the cause of irreversibility in shock *In* Mills L J, and Moyer, J H (eds) Shock and Hypotension Pathogenesis and Treatment Grune & Stratton, 1965, p 605

Crowell, J M, and Guyton, A C Evidence favoring a cardiac mechanism in irreversible hemorrhagic shock *Am J Physiol , 201* 893, 1961

Crowell J W, and Read W L *In vivo* coagulation–a probable cause of irreversible shock *Am J Physiol , 183* 565, 1955

Crowell, J W, and Smith, E E Oxygen deficit and irreversible hemorrhagic shock *Am J Physiol , 206* 313, 1964

Crowell J W, *et al* The mechanism of death after resuscitation following acute circulatory failure. *Surgery, 38* 696, 1955

Crowell, J W, *et al* The effect of varying the hematocrit ratio on the susceptibility to hemorrhagic shock *Am J Physiol , 192* 171, 1958

Crowell, J W, *et al* Effect of ellopurinol on hemorrhagic shock *Am. J Physiol , 216* 744, 1969

Filkins, J P Microkinins and the metabolic pathophysiology of septic shock *Fed Proc , 44* 300, 1985

Franklin, J, and Doelp, A Shocktrauma New York, St. Martin's Press, 1980

Guyton, A C, and Crowell, J W. Dynamics of the heart in shock. *Fed. Proc 20* 51, 1961

Guyton, A C, *et al* Cardiac Output and Its Regulation. Philadelphia, W B Saunders Co , 1973

Intensive Care of the Surgical Cardiopulmonary Patient. Chicago, Year Book Medical Publishers, 1983

Jameson, G A, and Greenwalt, T J (eds.) Blood Substitutes and Plasma Expanders New York, A R. Liss, 1978

Jones C E, *et al* A cause-effect relationship between oxygen deficit and irreversible hemorrhagic shock *Surgery, 127* 93, 1968

Jones, C E, *et al* Significance of increased blood uric acid following extensive hemorrhage *Am J Physiol , 214* 1374, 1968

Jones, C E, *et al* Cardiac output and physiological mechanisms in circulatory shock *In* MTP International Review of Science Physiology Vol 1 Baltimore University Park Press, 1974, p 233

Kovach, A G B, and Sandor, P Cerebral blood flow and brain function during hypotension. *Annu Rev Physiol , 38*:571, 1976

Lefer, A M. Eicosanoids as mediators of ischemia and shock *Fed Proc , 44* 276, 1985

Lefer, A M, and Martin, J Mechanism of the protective effect of corticosteroids in hemorrhagic shock. *Am J Physiol , 216* 314, 1969

Lefer, A M, and Martin, J Relationship of plasma peptides to the myocardial depressant factor in hemorrhagic shock in cats. *Circ. Res , 26* 59, 1970

Lefer, A M, *et al* Characterization of a myocardial depressant factor present in hemorrhagic shock *Am J Physiol., 213* 492, 1967.

McCord, J M Oxygen-derived free radicals in postischemic tissue injury *N Engl J. Med , 312* 159, 1985

Movat, H Z Microcirculation in disseminated intravascular coagulation induced by endotoxins *In* Renkin, E M, and Michel, C C (eds.) Handbook of Physiology, Sec 2, Vol IV Bethesda, American Physiological Society, 1984, p. 1047.

Rothe, C F, *et al* Control of total vascular resistance in hemorrhagic shock in the dog *Circ Res , 12* 667. 1963

Schmid Schonbein, H, and Textel, P (eds) Basic Aspects of Blood Trauma. Hingham, Mass, Kluwer Boston, 1979

Smith, E E, and Crowell, J. W Effect of hemorrhagic hypotension on oxygen consumption of dogs *Am J Physiol , 207* 647, 1964

Weil, M H, and Shubin, H· Shock Baltimore, Williams & Wilkins, 1969

Winkelman, J L· Essentials of Basic Life Support Minneapolis, Burgess, 1981

29

Muscle Blood Flow During Exercise; Cerebral, Splanchnic, and Skin Blood Flows

The blood flow in many special areas of the body, such as the lungs and the heart, has already been discussed in previous chapters, and the circulation in the kidney, in the uterus, and in the fetus will be discussed later in the text. In the present chapter the characteristics of blood flow in some of the other important vascular areas, such as the muscles, the brain, the splanchnic system, and the skin, are presented.

BLOOD FLOW THROUGH SKELETAL MUSCLES AND ITS REGULATION IN EXERCISE

Very strenuous exercise is the most stressful condition that the normal circulatory system faces. This is true because the blood flow in muscles can increase more than 20-fold (a greater increase than in any other tissue of the body) and also because there is such a very large mass of skeletal muscle in the body. The product of these two factors is so great that the total muscle blood flow in the normal young adult can increase during heavy exercise from the normal level of less than 1 liter per minute to as great as 20 liters per minute, high enough to increase the cardiac output to as much as five times normal and in the well-trained athlete to as much as six to seven times normal.

RATE OF BLOOD FLOW THROUGH THE MUSCLES

During rest, blood flow through skeletal muscle averages 3 to 4 ml per minute per 100 grams of muscle. However, during extreme exercise this rate can increase as much as 15- to 25-fold, rising to 50 to 80 ml per 100 g of muscle.

Intermittent Flow During Muscle Contraction. Figure 29–1 illustrates a study of blood flow changes in the calf muscles of the human leg during strong rhythmic muscular exercise. Note that the flow increases and decreases with each muscle contraction, decreasing during the contraction phase and increasing between contractions. At the end of the rhythmic contractions the blood flow remains very high for a few seconds longer but then gradually fades toward normal during the next few minutes.

The cause of the decreased flow during muscle contraction is compression of the blood vessels by the contracted muscle. During strong *tetanic* contraction, which causes sustained compression of the blood vessels, the blood flow can be almost totally stopped.

Opening of Muscle Capillaries During Exercise. During rest, only 20 to 25 per cent of the muscle capillaries are open. But during strenuous exercise all the capillaries open up, which can be demonstrated by studying histologic specimens removed from muscle during exercise. This opening up of dormant capillaries allows most of the increased blood flow. It also diminishes the distance that oxygen and other nutrients must diffuse from the capillaries to the muscle fibers and contributes a much increased surface area through which nutrients can diffuse from the blood.

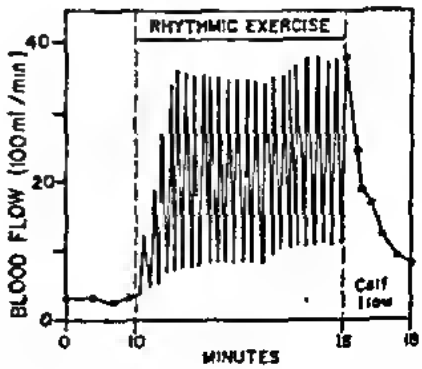


Figure 29–1. Effects of muscle exercise on blood flow in the calf of a leg during strong rhythmic contraction. The blood flow was much less during contraction than between contractions. (From Barcroft and Dornhorst: *J. Physiol.*, 109:402, 1949.)

CONTROL OF BLOOD FLOW THROUGH THE SKELETAL MUSCLES

Local Regulation. The tremendous increase in muscle blood flow that occurs during skeletal muscle activity is caused primarily by local effects in the muscles acting directly on the arterioles to cause vasodilatation.

This local increase in blood flow during muscle contraction is probably caused by several different factors all operating at the same time. One of the most important of these is reduction of oxygen in the muscle tissues. That is, during muscle activity the muscle utilizes oxygen very rapidly, thereby decreasing the oxygen concentration in the tissue fluids This in turn causes vasodilatation either because the vessel walls cannot maintain contraction in the absence of oxygen or because oxygen deficiency causes release of vasodilator substances The vasodilator substance that has been suggested most widely in recent years has been adenosine, but recent experiments have shown that adenosine cannot cause long-sustained vasodilation in skeletal muscle Furthermore, even after the muscle blood vessels have become insensitive to the vasodilator effects of adenosine, these same vessels will dilate fully in response to muscle activity.

Other vasodilator substances released during muscle contraction include potassium ions, acetylcholine, adenosine triphosphate, lactic acid, and carbon dioxide. Unfortunately, we still do not know quantitatively how great a role each of these plays in increasing muscle blood flow during muscle activity; this subject was discussed in more detail in Chapter 20

Nervous Control of Muscle Blood Flow. In addition to the local tissue regulatory mechanism, the skeletal muscles are also provided with sympathetic vasoconstrictor nerves and, in some species of animals, sympathetic vasodilator nerves as well

Sympathetic Vasoconstrictor Nerves. The sympathetic vasoconstrictor nerve fibers secrete norepinephrine and when maximally stimulated can decrease blood flow through the muscles to perhaps one-half to one-fourth normal. This represents rather poor vasoconstriction in comparison with that caused by sympathetic nerves in some other areas of the body in which blood flow can be almost completely blocked Yet even this degree of vasoconstriction is of physiological importance in circulatory shock and during other periods of stress when it is desirable to reduce blood flow through the many muscles of the body

In addition to the norepinephrine secreted at the sympathetic vasoconstrictor nerve endings, the adrenal medullae secrete large amounts of additional norepinephrine and epinephrine into the circulating blood during strenuous exercise. The circulating norepinephrine acts on the muscle vessels to cause a vasoconstrictor effect similar to that caused by direct sympathetic nerve stimulation. The epinephrine, on the other hand, has a vasodilator effect because epinephrine excites the *beta* receptors of the vessels, which are vasodilator receptors, in contrast to the *alpha* vasoconstrictor receptors excited by the norepinephrine. These receptors are discussed in Chapter 57.

Sympathetic Vasodilator Fibers. In the cat and some other lower animals there are also sympathetic vasodilator fibers that secrete acetylcholine—this hormone, in turn, causing the vasodilatation. However, such fibers have not yet been proved to occur in the human being. Instead, as noted previously, circulating epinephrine from the adrenal medullae acting at beta receptors in the muscle arterioles causes vasodilation, and this may subserve the same function as the vasodilator system of the lower animals

In the cat, maximal stimulation of the sympathetic vasodilator fibers to the skeletal muscles can increase blood flow by 400 per cent. These fibers are activated by a special nervous pathway beginning in the cerebral cortex, in close association with the motor areas for control of muscular activity, and passing downward through the hypothalamus and brain stem to the spinal cord When the motor cortex initiates muscle activity, it simultaneously excites the vasodilator fibers to the active muscles, and vasodilatation occurs immediately, several seconds before the local vasodilatory effects can take place as a result of the muscle activity itself Thus, it seems that this vasodilator system has the important *function of initiating extra blood flow through the muscles at the onset of muscular activity.* However, the vasodilator fibers are probably of little or no importance in maintaining increased blood flow during prolonged muscular activity, for both in lower animals and in human beings the final degree of vasodilatation that occurs during exercise is not discernibly different when the muscle is normally innervated with sympathetics or denervated.

CIRCULATORY READJUSTMENTS DURING EXERCISE

Three major effects occur during exercise that are essential for the circulatory system to supply the tremendous blood flow required by the muscles. These effects are (1) mass discharge of the sympathetic nervous system throughout the body with consequent stimulatory effects on the circulation, (2) increase in cardiac output, and (3) increase in arterial pressure.

Mass Sympathetic Discharge. At the onset of exercise, signals are transmitted not only from the brain to the muscle to cause muscle contraction but also from the higher levels of the brain into the vasomotor center to initiate mass sympathetic discharge Simultaneously, the parasympathetic signals to the heart are greatly attenuated. Therefore, two major circulatory effects result First, the heart is stimulated to greatly increased heart rate and pumping strength as a result of the sympathetic drive to the heart, as well as release of the heart from the normal parasympathetic inhibition. Second, all the blood vessels of the peripheral circulation are strongly contracted except the vessels in the active muscles, which are strongly vasodilated by the local vasodilator effects in the muscles themselves. Thus, the heart is stimulated to supply the increased blood flow required by the muscles, and blood flow through most nonmuscular areas of the body is temporarily reduced, thereby temporarily "lending" their blood supply to the muscles. This effect accounts for as much as two liters of extra blood flow to the muscles, which is exceedingly important when one thinks of the wild animal running for its life, because even a fractional increase in running speed may make the difference between life and death. However, two of the organ circulatory systems, the coronary and cerebral systems, are spared this vasoconstrictor effect because both these circulatory areas have very poor vasoconstrictor innervation—fortunately so because both the heart and the brain are as essential to exercise as are the skeletal muscles themselves.

A Muscle Reflex That Stimulates the Sympathetic Nervous System. Aside from the sympathetic stimulation caused by direct signals from the brain, reflex signals from the contracting muscles are also believed to pass up the spinal cord to the vasomotor center and to excite the sympathetic nerves These signals have been postulated to be initiated by metabolic end-products acting on small sensory nerve endings in the muscle tissue

Increase in Cardiac Output. The increase in cardiac output that occurs during exercise results mainly from the intense, locally induced vasodilatation in active muscles As was explained in Chapter 23 in relation to the basic theory of cardiac output regulation, local vasodilatation increases the venous return of blood back to the heart The heart in turn pumps this extra returning blood and sends it immediately back to the muscles through the arteries Thus, it is mainly the muscles themselves that determine the amount of increase in cardiac output—up to the limit of the heart's ability to respond

However, another factor that greatly helps cause the large increase in venous return is the *strong sympathetic stimulation of the veins.* This stimulation greatly increases the mean systemic filling pressure, sometimes to as high as 30 mm Hg (four times normal), and is therefore important in increasing venous return

Mechanisms by Which the Heart Increases Its Output. One of the principal mechanisms by which the heart increases its output during exercise is the Frank-Starling mechanism, which was discussed in Chapter 13 Via this mechanism, when increased quantities of blood flow from the veins into the heart and dilate its chambers, the heart muscle contracts with increased force, thus also pumping an increased volume of blood with each heart beat However, in addition to this basic intrinsic cardiac mechanism, the heart is also strongly stimulated by the sympathetic nervous system, and the normal parasympathetic inhibition is reduced or eliminated The net effects are greatly increased heart rate (occasionally to as high as 200 beats per minute) and almost doubling of the cardiac muscle strength of contraction These two effects combine to make the heart capable of pumping at least 100 per cent more blood than would be true based on the Frank-Starling mechanism alone

Increase in Arterial Pressure—Importance of Increased Sympathetic Activity. The mass sympathetic discharge throughout the body during exercise and the resultant vasoconstriction of most of the blood vessels besides those in the active muscles almost always increase the arterial pressure during exercise. This increase can be as little as 20 or as great as 80 mm Hg, depending on the conditions under which the exercise is performed When a person performs exercise under very tense conditions but uses only a few muscles, the sympathetic response occurs everywhere in the body but vasodilatation occurs in only a few muscles. Therefore, the net effect is mainly one of vasoconstriction, often increasing the mean arterial pressure to as high as 180 mm Hg Such a condition occurs in a person standing on a ladder and nailing with a hammer on the ceiling The tenseness of the situation is obvious, and yet ant of muscle vasodilatation is relatively slight. n the other hand, when a person performs wholely exercise, such as running or swimming, the Increase in arterial pressure is often only 20 to 40 mm Hg. The lack of a tremendous rise in pressure results from the extreme vasodilatation occurring in large masses of muscle.

In rare instances, persons are found in whom the sympathetic nervous system is absent, either congenitally or because of surgical removal. When such a person exercises, instead of the arterial pressure rising the pressure actually falls—sometimes to as low as one-half normal, and the cardiac output rises only about one third as much as it does normally. Therefore, one can readily understand the major importance of increased sympathetic activity during exercise.

Importance of the Arterial Pressure Rise During Exercise. In the well-trained athlete the muscle blood flow can increase at least 20-fold Though most of this increase results from vasodilatation in the active muscles, the increase in arterial pressure also plays an important role If one remembers that an increase in pressure not only forces extra blood through the muscle because of the increased pressure itself but also dilates the blood vessels, it can be seen that as little as a 20 to 40 mm Hg rise in pressure can at times actually double peripheral blood flow

It is especially important to note that in animals or human beings who do not have a sympathetic nervous system, the fall in arterial pressure that occurs during exercise has a strong negating effect on the rise in cardiac output that normally occurs. In such instances, the cardiac output can almost never be increased more than twofold, instead of the four- to sevenfold that can occur when the arterial pressure rises above normal

The Cardiovascular System as the Limiting Factor in Heavy Exercise. The capability of an athlete to enhance cardiac output and consequently to deliver increased quantities of oxygen and other nutrients to his or her tissues is the major factor that determines the degree of prolonged heavy exercise that the athlete can sustain For instance, the speed of a marathon runner is almost directly proportional to the ability to enhance cardiac output. Therefore, the ability of the circulatory system to adapt to exercise is equally as important as the muscles themselves in setting the limit for the performance of muscle work.

THE CEREBRAL CIRCULATION

NORMAL RATE OF CEREBRAL BLOOD FLOW

The normal blood flow through the brain tissue of the adult averages 50 to 55 ml per 100 grams of brain per minute For the entire brain of the average adult, this is approximately 750 ml per minute, or 15 per cent of the total resting cardiac output

REGULATION OF CEREBRAL BLOOD FLOW

As in most other vascular areas of the body, cerebral blood flow is highly related to the metabolism of the cerebral tissue. At least three different metabolic factors have been shown to have potent effects on cerebral blood flow. These are carbon dioxide concentration, hydrogen ion concentration, and oxygen concentration. An *increase* in either the carbon dioxide or the hydrogen ion concentration increases cerebral blood flow, whereas a *decrease* in oxygen concentration increases the flow.

Regulation of Cerebrul Blood Flow In Response to Excess Carbon Dioxide or Hydrogen Ion Concentration. An increase in carbon dioxide concentration in the arterial blood perfusing the brain greatly increases cerebral blood flow. This is illustrated in Figure 29–2, which shows that *doubling the arterial* P_{CO_2} by breathing carbon dioxide also approximately doubles the blood flow.

Carbon dioxide increases cerebral blood flow by combining with water in the body fluids to form carbonic acid, with subsequent dissociation to form hydrogen ions. The hydrogen ions then cause vasodilatation of the cerebral vessels—the dilatation being almost directly proportional to the increase in hydrogen ion concentration

Any other substance that increases the acidity of the brain tissue, and therefore also increases the hydrogen ion concentration, increases blood flow as well Such substances include lactic acid, pyruvic acid, or any other acidic material formed during the course of metabolism

Importance of the Carbon Dioxide and Hydrogen Control of Cerebral Blood Flow. Increased hydrogen ion concentration greatly depresses neuronal activity Therefore, it is fortunato that an increase in hydrogen ion concentration causes an increase in the blood flow, which In turn carries both carbon dioxide and other acidic substances away from the brain tissues Loss of the carbon dioxide removes carbonic acid from the tissues, and this, along with the removal of other acids, reduces the hydrogen ion concentration back toward normal. Thus, this mechanism helps maintain a constant hydrogen ion concentration in the cerebral fluids and therefore also maintains a normal level of neuronal activity.

Oxygen Deficiency as a Regulator of Cerebral Blood Flow. Except during periods of intense brain activity, the utilization of oxygen by the brain tissue remains within very narrow limits—within a few per cent of 3.5 ml of oxygen per 100 grams of brain tissue per minute. If the blood flow to the brain ever becomes insufficient to supply this needed amount of oxygen, the oxygen deficiency mechanism for vasodilatation, which was discussed in Chapter 20 and which functions in essentially all tissues of the body, immediately causes vasodilatation, returning the blood flow and transport of oxygen to the cerebral tissues to near normal Thus, this local blood flow regulatory mechanism is much the same as that existing in the coronary and skeletal muscle circulations and in many other circulatory areas of the body

Experiments have shown that a decrease in cerebral *tissue* P_{O_2} below approximately 30 mm Hg (normal value *is 35 to 40 mm Hg*) will begin to increase cerebral blood flow. This is very fortuitous because brain function becomes deranged at not much lower values of P_{O_2}, especially so at P_{O_2} levels below 20 mm Hg. Even coma can result at these low levels. Thus, the oxygen mechanism for local regulation of cerebral blood flow is a very important protective response against diminished cerebral neuronal activity and, therefore, against derangement of mental capability.

Measurement of Cerebral Blood Flow and Effect of Cerebral Activity on the Flow. A method has recently been developed to record blood flow in as many as 256 isolated segments of the human cerebral cortex simultaneously. A radioactive substance, usually radioactive xenon, is injected into the carotid artery; then the radioactivity of the cortex as the radioactiva substance passes through the brain tissue is recorded. Two hundred and fifty-six small radioactive scintillation detectors are focused on the same number of separate parts of the cortex; the rate of decay of the radioactivity after it once appears in each tissue segment is a direct measure of the rate of blood flow through the segment.

Using this technique, it has become clear that the blood flow in each individual segment of the brain changes within seconds in response to changes in local neuronal activity. For instance, simply clasping the hand causes an immediate increase in blood flow in the motor cortex of the opposite side of the brain. Reading a book increases the blood flow in multiple areas of the brain, especially in the occipital cortex and in the language areas of the temporal cortex This measuring procedure can also be used to localize the origin of epileptic attacks, for the blood flow increases acutely and markedly (as much as sevenfold!) in the focal point of the attack at its very onset

Illustrating the effect of local neuronal activity on cerebral blood flow, Figure 29–3 shows an increase in occipital blood flow recorded in a cat when intense light was shone into its eyes for a period of 0 5 minute.

Autoregulation of Cerebral Blood Flow When the Arterial Pressure Is Changed. Cerebral blood flow is autoregulated extremely well between the pressure limits of 60 and 140 mm Hg That is, the arterial pressure can be decreased acutely to as low as 60 mm Hg or increased to as high as 140 mm Hg without a significant change in cerebral blood flow. In persons who have *hypertension, this autoregulatory range shifts to even*

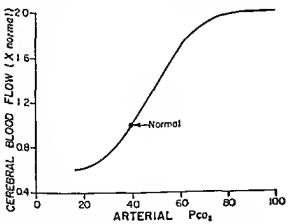


Figure 29–2. Relationship between arterial P_{CO_2} and cerebral blood flow

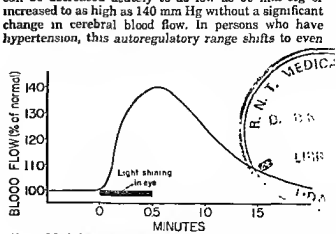


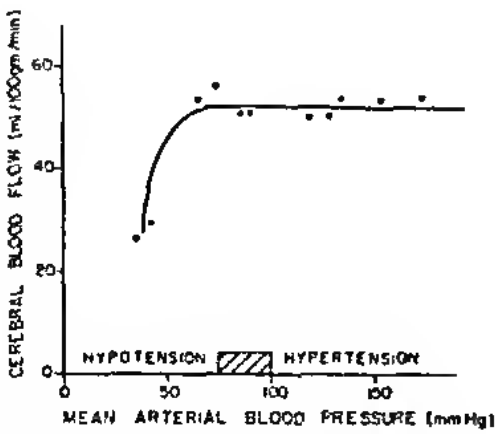
Figure 29–3. Increase in blood flow to the occipital regions of the brain when a light is flashed in the eyes of an animal.

higher pressure levels, up to as high as 180 to 200 mm Hg. This effect is illustrated in Figure 29–4, which shows cerebral blood flows measured both in normal human beings and in hypertensive patients. Note the extreme constancy of cerebral blood flow between the limits of 60 and 180 mm Hg mean arterial pressure. On the other hand, if the arterial pressure does fall below 60 mm Hg, cerebral blood flow then becomes severely compromised, and if the pressure rises above the upper limit of autoregulation, the blood flow rises rapidly and can cause severe overstretch of the cerebral blood vessels, sometimes resulting in serious brain edema.

Role of the Sympathetic Nervous System in Regulating Cerebral Blood Flow. The cerebral circulatory system has a strong sympathetic innervation that passes upward from the superior cervical sympathetic ganglia along with the cerebral arteries. This innervation supplies both the large superficial arteries and the small arteries that penetrate into the substance of the brain. However, neither transection of these sympathetic nerves nor stimulation of them normally causes significant change in the cerebral blood flow. Therefore, it was long stated that the sympathetic nerves play essentially no role in regulating cerebral blood flow.

However, recent experiments have shown that sympathetic stimulation can, under some conditions, constrict the cerebral arteries markedly. The reason that this usually does not occur is that the local blood flow autoregulatory mechanism is so powerful that it normally compensates almost entirely for the effects of the sympathetic stimulation. Yet, in those conditions in which the autoregulatory mechanism fails to compensate enough, then sympathetic control of cerebral blood flow becomes quite important. For instance, when the arterial pressure rises to a very high level during strenuous exercise and during other states of excessive circulatory activity, the sympathetic nervous system constricts the large and intermediate-sized arteries and prevents the very high pressure from ever reaching the smaller blood vessels. Experiments have shown that this is important in preventing the occurrence of a vascular hemorrhage into the brain—that is, for preventing the occurrence of cerebral stroke.

Also, sympathetic reflexes are believed to cause vasospasm in the intermediate and large arteries in some instances of brain damage, such as after a cerebral stroke has occurred or in patients with a subdural hematoma or brain tumor.



29–4. Relationship of mean arterial pressure to cerebral blood flow in normotensive, hypotensive, and hypertensive persons. (Modified from Lassen: *Physiol. Rev.*, 39:183, 1959.)

Brain Edema

One of the most serious complications of abnormal cerebral hemodynamics is the development of brain edema. Because the brain is encased in a solid vault, the accumulation of edema fluid compresses the blood vessels, with eventual depression of blood flow and destruction of brain tissue.

The usual cause of brain edema is either greatly increased capillary pressure or damage to the capillary endothelium. One cause of excessively high capillary pressure is a sudden increase in the cerebral blood pressure to levels too high for the autoregulatory mechanism to cope with. However, the most common cause is brain concussion, in which the brain tissues and capillaries are traumatized and capillary fluid leaks into the traumatized tissues. Once brain edema begins, it often initiates a vicious cycle because of the following positive feedback. The edema compresses the vasculature. This in turn decreases the blood flow and causes brain ischemia. The ischemia causes arteriolar dilatation with increased capillary pressure. The increased capillary pressure then causes more edema fluid, so that the edema becomes progressively worse. Once this vicious cycle has begun, heroic measures must be used to prevent total destruction of the brain. One such measure is to infuse intravenously a concentrated osmotic substance, such as a very concentrated mannitol solution. This pulls fluid by osmosis from the brain tissue and breaks up the vicious cycle. Another procedure is to remove fluid quickly from the lateral ventricles of the brain via ventricular puncture, thereby relieving the intracerebral pressure.

THE SPLANCHNIC CIRCULATION

A large share of the cardiac output flows through the vessels of the intestines and through the spleen, finally coursing into the portal venous system and then through the liver, as illustrated in Figure 29–5. This is called the *portal circulatory system*, and it, plus the arterial blood flow into the liver, is called the *splanchnic circulation*.

BLOOD FLOW THROUGH THE LIVER

About 1100 ml of portal blood enter the liver each minute. This flows through the *hepatic sinuses* in close contact with the cords of liver parenchymal cells. Then it enters the *central venules* of the liver lobules and from there flows into the vena cava.

In addition to the portal blood flow, approximately 350 ml of blood flow into the liver each minute through the hepatic artery, making a total hepatic flow of about 1450 ml per minute, or an average of 29 per cent of the total cardiac output. The hepatic arterial blood flow, in addition to supplying about one quarter of the blood flowing through the liver sinuses along with the portal blood, also supplies nutrition to the connective tissue and especially to the walls of the bile ducts. Therefore, loss of hepatic arterial flow can be lethal because this often causes necrosis of the basic liver structures.

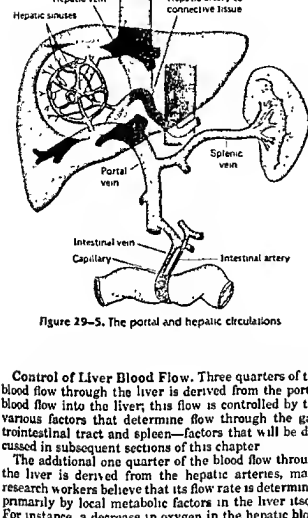


Figure 29–5. The portal and hepatic circulations

Control of Liver Blood Flow. Three quarters of the blood flow through the liver is derived from the portal blood flow into the liver; this flow is controlled by the various factors that determine flow through the gastrointestinal tract and spleen—factors that will be discussed in subsequent sections of this chapter

The additional one quarter of the blood flow through the liver is derived from the hepatic arteries, many research workers believe that its flow rate is determined primarily by local metabolic factors in the liver itself For instance, a decrease in oxygen in the hepatic blood causes an increase in hepatic arterial blood flow However, other experiments have shown that a decrease in portal blood flow into the liver is compensated by increase in hepatic artery flow, thus maintaining total hepatic flow at a very constant level This is important because of the many metabolic functions that need to continue at constant rates regardless of what happens to portal blood flow.

Reservoir Function of the Liver. Because the liver is an expandable and compressible organ, large quantities of blood can be stored in its blood vessels Its normal blood volume, including both that in the hepatic veins and hepatic sinuses, is 450 ml, or almost 10 per cent of the total blood volume However, when high pressure in the right atrium causes back pressure in the liver, the liver expands, and as much as 1 liter of extra blood occasionally is thereby stored in the hepatic veins and sinuses. This occurs especially in cardiac failure with peripheral congestion, which was discussed in Chapter 26.

Thus, in effect, the liver is a large, expandable, venous organ capable of acting as a valuable blood reservoir in times of excess blood volume and capable of supplying extra blood in times of diminished blood volume

Control of Blood Storage in the Liver by the Sympathetic Nervous System. The storage portions of the liver, particularly the large veins and much less so the sinusoids, are vasoconstricted by sympathetic stimulation. Therefore, in times of circulatory stress when the sympathetic nervous system is discharging strongly, a large portion of blood stored in the liver is expelled into the general circulation within one to four minutes. In the normal human being this can probably amount to as much as 300 ml of blood. Thus, the liver is one of the most important sources of blood to fill the other portions of the circulation in times of need, especially during heavy exercise or after severe hemorrhage.

Permeability of the Hepatic Sinuses. The hepatic sinuses are lined with an endothelium similar to that of the capillaries, but its permeability is extreme in comparison with that of usual capillaries—so much so that even the proteins of the blood diffuse into the extravascular spaces of the liver almost as easily as fluids. Fortunately, the pressure in the sinuses is only 6 to 8 mm Hg so that most of the proteins that diffuse out of the sinuses also diffuse back in. Yet, the remainder passes into the liver lymphatics, so that liver lymph normally contains almost as great a protein concentration as that of the plasma itself. Also, the lymph from the liver accounts for one half to two thirds of the total body lymph flow

This extreme permeability of the liver sinuses brings the fluids of the hepatic blood into extremely close contact with the liver parenchymal cells, thus facilitating rapid exchange of nutrient materials between the blood and the liver cells

The Blood-Cleansing Function of the Liver. Blood flowing through the intestinal capillaries picks up many bacteria from the intestines Indeed, a sample of blood from the portal system almost always grows colon bacilli when cultured, whereas growth of colon bacilli from blood in the systemic circulation is extremely rare Special high-speed motion pictures of the action of Kupffer cells, the large phagocytic reticuloendothelial cells that line the hepatic sinuses, have demonstrated that these cells can cleanse blood extremely efficiently as it passes through the sinuses; when a bacterium comes into momentary contact with a Kupffer cell, in less than 0 01 second the bacterium passes inward through the wall of the Kupffer cell to become permanently lodged therein until it is digested Probably not over 1 per cent of the bacteria entering the portal blood from the intestines succeeds in passing through the liver into the systemic circulation

Measurement of Blood Flow Through the Liver in Man. Several indirect methods for measuring blood flow through the liver, similar to those used for measuring blood flow through the heart and brain, have been developed The most important of these is the following One of the many substances removed from the blood by the liver is the dye *indocyanine green* To measure the liver blood flow, this dye is infused into a vein at a continuous rate until its concentration becomes steady in the circulating blood Then a sample of blood is removed from a peripheral vein while at the same time another sample of blood is removed from a catheter previously placed into the hepatic vein under fluoroscopic control. Because the liver removes much of the dye as the blood passes through it, there is a concentration difference between the systemic blood and the liver venous blood. Also, because the concentration in the systemic blood is constant, the rate of removal of the dye by the liver is equal to the rate of injection of dye Consequently, the liver blood flow can be calculated

using the *Fick equation* for flow that was explained for measurement of cardiac output in Chapter 23. That is:

$$\text{Liver blood flow} = \frac{\text{Rate of infusion of dye}}{\text{Concentration difference}}$$

BLOOD FLOW THROUGH THE INTESTINAL VESSELS

About four fifths of the portal blood flow originates in the intestines and stomach (about 850 ml per minute), and the remaining one fifth originates in the spleen and pancreas. Over two thirds of the intestinal flow is to the mucosa to supply the energy needed for forming the intestinal secretions and for absorbing the digested food.

Control of Gastrointestinal Blood Flow. Blood flow in the gastrointestinal tract is controlled in almost exactly the same way as in most other areas of the body mainly by local regulatory mechanisms. Furthermore, blood flow to the mucosa and submucosa, where the glands are located and where absorption occurs, is controlled separately from blood flow to the musculature. When glandular secretion increases, so does mucosal and submucosal blood flow—sometimes as much as eightfold! Likewise, when motor activity of the gut increases, blood flow in the muscle layers increases.

However, the precise mechanisms by which alterations in gastrointestinal activity alter the blood flow are not completely understood. It is known that decreased availability of oxygen to the gut increases local blood flow in the same way that this occurs elsewhere in the body; therefore local regulation of blood flow in the gut might occur almost entirely secondarily to changes in metabolic rate. On the other hand, it is also known that various peptide hormones are released from the mucosa of the intestinal tract during the digestive process and that some of these also cause mucosal vasodilatation. The best known of these hormones are *vasoactive intestinal peptide, gastrin, secretin,* and *cholecystokinin*. Also, it has been claimed that some or all of the gastrointestinal glands form two *kinins*, the substances *kallidin* and *bradykinin*, at the same time that they release their secretions. These kinins are powerful vasodilators and have been postulated to cause much of the mucosal vasodilatation. However, critical experiments have not yet proved this mechanism.

Nervous Control of Gastrointestinal Blood Flow. Stimulation of the parasympathetic nerves to the *stomach* and *lower colon* increases local blood flow at the same time that it increases glandular secretion. However, this increased flow probably results secondarily from the increased glandular activity and not as a direct effect of the nervous stimulation.

Sympathetic stimulation, in contrast, has a direct effect on essentially all the gastrointestinal tract by causing intense vasoconstriction of the arterioles with greatly decreased blood flow. However, after a few minutes of this vasoconstriction, the flow returns almost to normal via a mechanism called "autoregulatory escape." That is, the local metabolic vasodilator mechanisms that [illegible] elicited by ischemia become prepotent over the [illegible] vasoconstriction and, therefore, redilate the [illegible], thus causing return of the necessary nutrient blood flow to the gastrointestinal glands and muscle.

A major value of sympathetic vasoconstriction in the gut is that it allows the shutting off of splanchnic blood flow for short periods of time during heavy exercise when increased flow is needed by the skeletal muscle and heart.

Sympathetic stimulation causes especially strong vasoconstriction of the intestinal and mesenteric veins. Furthermore, this vasoconstriction does not "escape." Instead, it decreases the volume of these veins and thereby displaces large amounts of blood into other parts of the circulation. In hemorrhagic shock or other states of low blood volume, this mechanism can provide probably 200 to 300 milliliters of extra blood to sustain the general circulation.

PORTAL VENOUS PRESSURE

The liver offers a moderate amount of resistance to blood flow from the portal system to the vena cava. As a result, the pressure in the portal vein averages 8 to 10 mm Hg, which is considerably higher than the almost zero pressure in the vena cava. Because of this high portal venous pressure, the pressures in the intestinal venules and capillaries have a much greater tendency to become abnormally high than is true elsewhere in the body.

Blockage of the Portal System—"Cirrhosis" of the Liver. Frequently, extreme amounts of fibrous tissue develop within the liver structure, destroying many of the parenchymal cells and eventually contracting around the blood vessels, thereby greatly impeding the flow of portal blood through the liver. This disease process is known as *cirrhosis of the liver*. It results most frequently from alcoholism, but it can also follow ingestion of poisons such as carbon tetrachloride, virus diseases such as infectious hepatitis, or infectious processes in the bile ducts.

The portal system is also occasionally blocked by a large clot developing in the portal vein or in its major branches.

When the portal system is suddenly blocked, the return of blood from the intestines and spleen to the systemic circulation is tremendously impeded, the capillary pressure rising as much as 15 to 20 mm Hg, and the patient often dies within a few hours because of excessive loss of fluid from the capillaries into the lumina and walls of the intestines.

Development of Collateral Circulation from the Portal Venous System to the Systemic Veins. When portal blood flow is blocked slowly rather than suddenly, such as occurs in slowly developing cirrhosis of the liver, large collateral vessels develop between the portal veins and the systemic veins. The most important of these are collaterals from the splenic to the esophageal veins; these collaterals frequently become so large that they protrude deeply into the lumen of the esophagus and are then called *esophageal varicosities*. The esophageal mucosa overlying these varicosities eventually becomes eroded; and in many if not most of these patients, the erosion finally penetrates all the way into the varicosity and causes the person to bleed severely, often to bleed to death.

Ascites as a Result of Portal Obstruction. Ascites is free fluid in the peritoneal cavity. It results from exudation of fluid either from the surface of the liver or from the surfaces of the gut and its mesentery. Ascites usually will develop only in case outflow of blood from the liver into the inferior vena cava is blocked. This causes extremely high pressure in the liver sinusoids,

which in turn causes fluid to weep from the surfaces of the liver. The weeping fluid is almost pure plasma, containing tremendous quantities of protein The protein, because it causes a high colloid osmotic pressure in the abdominal fluid, then pulls by osmosis additional fluid from the surfaces of the gut and mesentery

On the other hand, obstruction of the portal vein, without directly involving the liver, rarely causes ascites. If obstruction occurs acutely, the person is likely to die of shock within hours because of fluid loss into the gut; if it occurs slowly, collateral vessels can usually develop enough to prevent either the intestinal loss of fluid or the development of ascites

THE SPLENIC CIRCULATION

The Spleen as a Reservoir. The capsule of the spleen in many lower animals contains large amounts of smooth muscle, and sympathetic stimulation will cause intense contraction of the spleen Conversely, sympathetic inhibition results in considerable splenic expansion with consequent storage of blood

In humans, the splenic capsule is nonmuscular, but even so, dilatation of vessels within the spleen can still cause the spleen to store several hundred milliliters of blood at times, and then, under the influence of sympathetic stimulation, constriction of the vessels will express much of this blood into the general circulation Unfortunately, these effects are poorly understood, however, the spleen is so small, only 150 to 200 ml in volume, that this reservoir function in humans is of relatively little importance

As illustrated in Figure 29–6, two areas exist in the spleen for the storage of blood; the *venous sinuses* and the *pulp*. Small vessels flow directly into the venous sinuses, and when the spleen distends, the venous sinuses swell, thus storing blood

In the splenic pulp, the capillaries are very permeable, so that much of the blood passes first into the pulp and then oozes through this before entering the venous sinuses. As the spleen enlarges, many cells become stored in the pulp. Therefore, the net quantity of red blood cells in the general circulation decreases slightly when the spleen enlarges The spleen probably can store enough cells that splenic contraction can probably cause the hematocrit of the systemic blood to increase in humans as much as 1 to 2 per cent and as much as 3 to 4 per cent in some lower animals

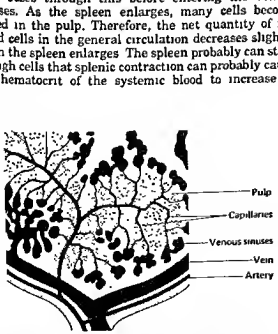


Figure 29–6. The functional structures of the spleen. (Modified from Bloom and Fawcett A Textbook of Histology 10th ed. Philadelphia, W. B. Saunders Company, 1975)

This ability of the spleen to store and release blood is important, at least in lower animals, in times of stress such as during strenuous exercise, for the circulatory system then needs an increased volume of blood so that rapid return of blood to the heart can be maintained At the same time, the blood that flows through the muscles needs to carry larger quantities of oxygen than normally. Therefore, the increased hematocrit of the circulating blood resulting from splenic contraction is also an aid during these periods of stress

The Blood-Cleansing Function of the Spleen—Removal of Old Cells. Blood passing through the splenic pulp before it enters the sinuses undergoes thorough squeezing Therefore, it is to be expected that fragile red blood cells would not withstand the trauma. For this reason, many of the red blood cells destroyed in the body have their final demise in the spleen After the cells rupture, the released hemoglobin and the cell stroma are ingested by the reticuloendothelial cells of the spleen.

Reticuloendothelial Cells of the Spleen. The pulp of the spleen contains many large phagocytic reticuloendothelial cells, and the venous sinuses are lined with similar cells These cells act as a cleansing system for the blood, similar to that in the venous sinuses of the liver When the blood is invaded by infectious agents, the reticuloendothelial cells of the spleen rapidly remove debris, bacteria, parasites, etc Also, in many infectious processes the spleen enlarges in the same manner that lymph glands enlarge and then performs its cleansing function even more adequately

Much of the spleen is filled with *white pulp*, which is in reality a large quantity of lymphocytes and plasma cells. These function in exactly the same way in the spleen as in the lymph glands to cause either humoral or lymphocytic immunity against toxins, bacteria, and so forth, as described in Chapter 6.

The Spleen as a Hemopoietic Organ. During fetal life, the splenic pulp produces blood cells in exactly the same manner that the red bone marrow in the adult produces cells. As the normal fetus approaches birth, the spleen normally loses this ability to produce cells, but, in some diseases, the spleen continues to produce cells even after birth. For instance, in the disease *erythoblastosis fetalis*, which results from excessive destruction of red blood cells by abnormal antibodies in the plasma, as discussed in Chapter 4, the fetus must produce ten or more times as many red blood cells as normally As a result, the hemopoietic function of the spleen persists even after birth.

CIRCULATION IN THE SKIN

PHYSIOLOGIC ANATOMY OF THE CUTANEOUS CIRCULATION

Circulation through the skin subserves two major functions: first, *nutrition of the skin tissues* and, second, *conduction of heat* from the internal structures of the body to the skin so that the heat can be removed from the body. To perform these two functions the circulatory apparatus of the skin is characterized by two major types of vessels, diagrammed in Figure 29–7: (1) the usual nutritive arteries, capillaries, and veins and (2)

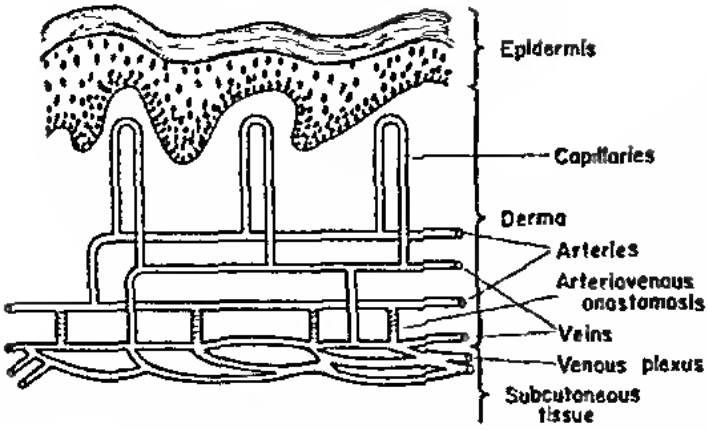


Figure 29-7. The skin circulation.

vascular structures concerned with heating the skin, consisting principally of (a) an extensive *subcutaneous venous plexus*, which holds large quantities of blood that can heat the surface of the skin, and (b) in some skin areas, *arteriovenous anastomoses*, which are large vascular communications directly between the arteries and the venous plexuses The walls of these anastomoses have strong muscular coats innervated by sympathetic vasoconstrictor nerve fibers that secrete norepinephrine. When constricted, they reduce the flow of blood into the venous plexuses to almost nothing, or when maximally dilated, they allow extremely rapid flow of warm blood into the plexuses The arteriovenous anastomoses are found principally in the volar surfaces of the hands and feet, the lips, the nose, and the ears, which are areas of the body most often exposed to maximal cooling In the other areas of the skin where arteriovenous anastomoses do not occur, large arterioles and capillaries serve a similar role in providing blood flow, though not as much, into the venous plexus

Rate of Blood Flow Through the Skin. The rate of blood flow through the skin is among the most variable of any part of the body, because the flow required to regulate body temperature changes markedly in response to, first, the rate of metabolic activity of the body, and, second, the temperature of the surroundings. This will be discussed in detail in Chapter 72 The blood flow required for nutrition is slight, so that this plays almost no role in controlling normal skin blood flow At ordinary skin temperatures, the amount of blood flowing through the skin vessels to provide heat regulation is about ten times as much as that needed to supply the nutritive needs of the tissues But, when the skin is exposed to extreme cold, the blood flow may then become so slight that nutrition begins to suffer—even to the extent, for instance, that the fingernails grow considerably more slowly in arctic climates than in temperate climates

Under ordinary cool conditions the blood flow to the skin is about 0 25 liter sq meter of body surface area, or a total of about 400 ml per minute in the average adult This can decrease to as little as 50 ml per minute in ere cold, and, when the skin is heated until maximal dilatation has resulted, the blood flow can be as uch as 2 to 3 liters per minute, thus illustrating both the extreme variability of skin blood flow and the great drain on cardiac output that can occur under hot conditions Indeed, many persons with borderline cardiac failure develop severe failure in hot weather because of the extra load on the heart and then revert from failur in cool weather

REGULATION OF BLOOD FLOW IN THE SKIN

Nervous Control of Cutaneous Blood Flow. Sinc most of the blood flow through the skin is to contr body temperature, and since this function in turn i regulated by the nervous system, the blood flow throug the skin is principally regulated by nervous mechanism rather than by local regulation, which is opposite to th regulation in most parts of the body

Temperature Control Center of the Hypothalamus. Located in the preoptic region of the anteric hypothalamus is a small center that is capable of cor trolling body temperature Heating this area cause vasodilation of essentially all the skin vessels of th body and also causes sweating. Cooling of the cente causes vasoconstriction and cessation of sweating I addition, temperature receptors in the skin also hel control the skin blood flow by causing reflex nerv signals to the skin blood vessels by way of the spin cord or all the way up to the hypothalamus and the back to the vessels The detailed function of this contr system will be discussed in Chapter 72 in relation t body temperature The important point in the preser discussion is that the hypothalamus controls blood flo in the skin in response to changes in body temperatu by two mechanisms (1) a vasoconstrictor mechanis and (2) a vasodilator mechanism.

The Vasoconstrictor Mechanism. The ski throughout the body is supplied with sympathetic va soconstrictor fibers that secreto norepinephrine at the endings This constrictor system is extremely powerfi in the feet, hands, lips, nose, and ears, which are th areas most frequently exposed to severe cold and whic are also the areas where large numbers of arteriovenou anastomoses are found At normal body temperature th sympathetic vasoconstrictor nerves keep these anast moses almost totally closed, but when the body becom overly heated the number of sympathetic impulses I greatly reduced so that the anastomoses dilate and allo large quantities of warm blood to flow into the venou plexuses, thereby promoting loss of heat from the bod

In the remainder of the body—that is, over the su faces of the arms, legs, and body trunk—almost n arteriovenous anastomoses are present, but neverthele vasoconstrictor control of the nutritive vessels can st effect major decreases in blood flow by constricting th local arterioles When the body becomes overheated, o the other hand, the sympathetic vasoconstrictor impu ses cease, and the blood flow to the skin vessels increase about twofold

Extreme Sensitivity of the Skin Blood Vessels t Circulating Norepinephrine and Epinephrine. I addition to the direct sympathetic vasoconstrictor effec in the skin, the skin blood vessels are also extremel sensitive to circulating norepinephrine and epinephrin Therefore, even in areas of the skin that might hav lost their sympathetic innervation, any mass discharg of the sympathetic nervous system will still cause in tense skin vasoconstriction. Sometimes the sensitivit of the skin to circulating norepinephrine and epinep rine is so great that the vasoconstriction can actuall

damage the skin. This response will be discussed in more detail later in the chapter in relation to *Raynaud's disease.*

The Vasodilator Mechanism. When the body temperature becomes excessive and sweating begins to occur, the blood flow through the skin of the forearms and trunk increases an additional severalfold, which occurs as a result of so-called "active" vasodilatation of the vessels. However, the basic mechanism by which this active dilation occurs is not completely known. This additional increase in blood flow does not occur except in the presence of sweating, and it does not occur in lower animals that do not have sweat glands. Therefore, it has been postulated that the sympathetic fibers that secrete acetylcholine to activate the sweat glands cause a secondary vasodilatation as follows. The increased activity of the sweat glands is postulated to cause these glands to release an enzyme called *kallikrein*, which in turn splits the polypeptide *bradykinin* from globulin in the interstitial fluids. Bradykinin in turn is a powerful vasodilator substance that could account for the greatly increased blood flow when sweating begins to occur. In opposition to this theory, however, is the fact that inhibition of the bradykinin mechanism does not completely block the increased blood flow associated with sweating. Therefore, other vasodilator peptides have also been suggested as the cause of the vasodilation.

Effect of Cold on the Skin Circulation. When cold is applied directly to the skin, the skin vessels constrict more and more down to a temperature of about 15 °C, at which point they reach their maximum degree of constriction. This constriction results primarily from increased sensitivity of the vessels to nerve stimulation, but it probably also results at least partly from a reflex that passes to the cord and then back to the vessels. At temperatures below 15° C the vessels begin to dilate. This dilation is caused by a direct local effect of the cold on the vessels themselves—probably paralysis of the contractile mechanism of the vessel wall or block of the nerve impulses coming to the vessels. In any event, at temperatures approaching 0° C the skin blood flow is sometimes considerably above normal. This vasodilatation in severe degrees of cold plays a purposeful role in preventing freezing of the exposed portions of the body, particularly the hands and ears.

Metabolic Regulation of Blood Flow in the Skin. Though metabolic regulation of blood flow usually plays only a small role in skin blood flow control, it does have an effect in those few instances in which skin blood flow becomes greatly decreased. For instance, if a person sits on the buttocks for 30 minutes or more and then stands, intense reddening of the affected skin will be noted. That is, the blood flow to the skin area has now increased to a marked extent, which is a typical *reactive hyperemia* response resulting from diminished availability of nutrients to the tissues during the period of compression. Thus, the local metabolic regulatory mechanism present in essentially all other tissues of the body is also present in the skin and can be called into play when needed to prevent nutritional damage to the tissues.

Shift of Blood from the Skin to the Remainder of the Circulation in Times of Circulatory Stress. The skin venous plexuses of the entire body, including those of the hands, feet, arms, legs, and trunk, are all strongly supplied with sympathetic vasoconstrictor innervation. In times of circulatory stress, such as during exercise, following severe hemorrhage, or even in states of anxiety, sympathetic stimulation of these venous plexuses can force large quantities of blood, estimated to be as much as 5 to 10 per cent of the total blood volume, into the internal vessels. Thus, the subcutaneous veins of the skin act as an important blood reservoir, often providing blood to serve other circulatory functions when needed.

COLOR OF THE SKIN IN RELATION TO SKIN TEMPERATURE

Much of the skin color is due to the color of the blood in the skin capillaries and veins. Therefore, when the skin is hot and arterial blood is flowing rapidly through these vessels, the skin is usually red. On the other hand, when the skin is cold and the blood is flowing extremely slowly, most of the oxygen is removed for nutritive purposes before the blood can leave the capillaries. Consequently, the capillaries and veins then contain large amounts of dark, deoxygenated blood that gives the skin a bluish hue.

Another cutaneous vascular effect that often affects skin color is severe constriction of the cutaneous vessels, as occurs when the sympathetic nervous system is strongly stimulated during circulatory shock. This expresses most of the blood out of the skin into other parts of the circulation. When this occurs the skin takes on the color of the subcutaneous connective tissue, which is composed principally of collagen fibers and has a whitish hue. Thus, in conditions in which the cutaneous vessels become greatly constricted, the skin has an ashen white pallor.

PHYSIOLOGY OF VASCULAR DISEASES OF THE LIMBS

RAYNAUD'S DISEASE

The local vasoconstrictor reflex that occurs normally during exposure to cold becomes especially marked in the limbs of some persons, particularly in many women who live in cold climates. This effect, called *Raynaud's disease,* occurs most frequently in the hands, though occasionally also in the feet. When the hands become even slightly cooled, they are likely to exhibit extreme vascular spasm—so much so that the hands do not receive sufficient flow of blood even for maintaining adequate nutrition. The tissues then develop *ischemic pain* in exactly the same manner that the heart develops pain in angina pectoris and as other tissues develop pain during ischemia. After the ischemia persists for several minutes, *hyperemia* often supervenes, causing dilatation of all the vessels of the hands so that the final stage is sometimes *intense reddening* of the hands. This effect is probably caused by the same factors that cause reactive hyperemia following temporary vascular occlusion in most other tissues, as discussed in Chapter 20. Strangely enough, the pain becomes even more severe at this moment, probably because of mobilization of the abnormal metabolic products that cause pain.

Though Raynaud's disease in its early stages is merely a physiologic entity that passes through the phases of blanching of the skin, pain, and reddening of the skin, in some patients it progresses over a period of years until the vessels become continuously constricted. In severe cases the blood flow can become decreased so much that the fingers become gangrenous.

Sympathectomy for Raynaud's Disease. To treat Raynaud's disease the sympathetic reflex is often blocked. This is done best by cutting the *preganglionic* fibers in the sympathetic chain at T-2 and T-3, which interrupts the sympathetic nerve impulses from the spinal cord to the hand The postganglionic fibers to the hand originate mainly in the stellate ganglion, but the stellate ganglion is not removed, for removal of postganglionic sympathetic fibers allows the blood vessels to become excessively sensitized to circulating norepinephrine and epinephrine. If this happens, the hands will still exhibit Raynaud's syndrome every time the adrenal glands are excited This phenomenon of sensitization will be discussed in Chapter 57 in connection with the various reactions of the sympathetic nervous system

PERIPHERAL ARTERIOSCLEROSIS

Arteriosclerosis, which causes intra-arterial plaques to develop, has already been discussed in connection with the coronary arteries This same process can occur throughout the arterial tree and not only in the coronaries Therefore, in old age many of the peripherol arteries will often have their intimal surfaces almost completely covered with lipid deposits or calcified plaques

Progressive obliteration of the peripheral arterial tree by arteriosclerosis results in all the symptoms that one would expect It causes pain in the muscles during activity and also in other tissues because of ischemia, and it is likely to cause gangrene and ulceration of the skin because of poor blood supply

A special difference between arteriosclerotic peripheral vascular disease and Raynaud's disease is that sympathectomy is not nearly so effective for treatment. This is true probably because the former is caused by obstruction of the large arteries rather than the small arteries, and the small arteries, which are affected most by the sympathetics, probably remain maximally vasodilated in severe arteriosclerosis whether the sympathetics are removed or not.

REFERENCES

Abramson, D I (ed) Circulatory Diseases of the Limbs A Primer New York, Grune & Stratton, 1978

American Physiological Society Peripheral Circulation and Organ Blood Flow Washington, D C , American Physiological Society, 1983

Anderson P O Vascular control in the colon and rectum *Scand J Gastroenterol* , 93 65 1984

Appenzeller, O , and Atkinson, R. (eds) Health Aspects of Endurance Training New York, S Karger, 1978

Apple, D F , Jr , and Contwell J D Medicine for Sport Chicago, Year Book Medical Publishers, 1979

Bes, A (ed) Proceedings of the Eleventh International Symposium on Cerebral Blood Flow and Metabolism. New York, Raven Press, 1983

Betz, F Cerebral blood flow Its measurement and regulation. *Physiol Rev* , *52* 595, 1972

Bissonnette, J M , et al Regulation of cerebral blood flow in the fetus. *J Dev Physiol* , *6* 275, 1984

Chapman, C B The physiology of exercise *Sci Am* , *212* 88 1965

[illegible] Sports Physiology Philadelphia, W B Saunders Co , 1979

Gelmers, H J : Calcium-channel blockers. Effects on cerebral blood flow and potential uses for acute stroke. *Am J Cardiol.*, *55* 144B, 1985

Gisvold, S E , and Steen, P A : Drug therapy in brain ischaemia *J Anaesthesiol* , 57 96, 1985

Granger, D N , and Kvietys, P R · The splanchnic circulation. Intra regulation. *Annu Rev Physiol* , *43* 409, 1981

Green, H D , and Kepchar, J H · Control of peripheral resistance major systemic vascular beds. *Physiol. Rev* , *39* 617, 1959

Greenleaf, J E Hyperthermia and exercise *In* Robertshaw, D (International Review of Physiology Environmental Physiology Vol 20 Baltimore, University Park Press, 1979, p 157.

Guth, P H Stomach blood flow and acid secretion *Annu Rev Physi* *44* 3, 1982

Guyton, A C , *et al.* Cardiac Output and Its Regulation. Philadelp W B Saunders Co , 1973

Harrison, M H Effects of thermal stress and exercise on blood vol in humans *Physiol. Rev* , *65* 149, 1985

Heistad, D D , and Kontos, H A Cerebral circulation. *In* Shepher T , and Abboud, F M (eds) Handbook of Physiology Sec. 2, Vol Bethesda, American Physiological Society, 1983, p 137.

Heros R C , and Zervas, N T Subarachnoid hemorrhage Ann *Med* , *34* 367, 1983

Jarrett, A Physiology and Pathophysiology of the Skin New Y Academic Press, 1984

Jensen, C R , and Fisher, G Scientific Basis of Athletic Conditio Philadelphia, Lea & Febiger, 1978

Juergens, J L., *et al* (eds.) Allen, Barker, Hines Peripheral Vasc Disease, 5th Ed Philadelphia, W B Saunders Co , 1980

Kontos, H A Regulation of the cerebral circulation *Annu Rev P* *iol.*, *43* 397, 1981

Kuschinsky, W , and Wahl, M Local chemical and neurogenic reg tion of cerebral vascular resistance *Physiol Rev* , *58* 656, 1978

Lassen, N A. Study of local cerebral blood flow *In* Dickinson, C and Marks J (eds.) Developments in Cardiovascular Medicine I caster, England, MTP Press, 1978 p 9

Lassen, N A , *et al* Brain function and blood flow *Sci Am* , *239* 4 1978

McDowall, D G Induced hypotension and brain ischaemia. *Br Anaesthesiol* , *57* 110, 1985

Mitchell J H , *et al* The exercise pressor reflex Its cardiovasc effects, afferent mechanisms, and central pathways. *Annu Rev P* *iol* , *45* 229, 1983

Moskalenko, Y E , *et al* Biophysical Aspects of Cerebral Circulat New York, Pergamon Press, 1979

Najarian, J S Advances in Vascular Surgery Chicago, Year B Medical Publishers, 1983

Pardridge, W M Brain metabolism A perspective from the blood b barrier *Physiol Rev* , *63* 1481, 1983

Rappaport, A M Hepatic blood flow Morphologic aspects and phy logic regulation *In* Javitt, N B (ed) International Review of Ph siology Liver and Biliary Tract Physiology I Vol. 21. Baltim University Park Press, 1980, p 1

Reed, G , and Devous, M Cerebral blood flow autoregulation hypertension *Am J Med Sci* , *289* 37, 1985

Reitz, B A Myocardial protection during cardiac surgery *Annu* [illegible] Med , 33 153, 1982

Roddie, I C Circulation to skin and adipose tissue *In* Shepherd, J and Abboud, F M (eds) Handbook of Physiology Sec. 2, Vol. Bethesda, American Physiological Society, 1983, p 285

Rowell, S , and Belfrage, E · Blood circulation in adipose tissue *Phy* *Rev* , *59* 1078, 1979

Shepherd, A P Local control of intestinal oxygenation and blood f *Annu Rev Physiol.*, *44* 13, 1982

Shepherd, J T Circulation to skeletal muscle *In* Shepherd, J T , Abboud F M (eds) Handbook of Physiology Sec. 2, Vol Bethesda, American Physiological Society, 1983, p 319

Smith, C R , and Rodeheffer, R J , Raynaud a phenomenon. Pathop iologic features and treatment with calcium-channel blockers. *Am Cardiol* , *55* 154B, 1985

Strauss, R H (ed.) Sports Medicine and Physiology Philadelp W B Saunders Co , 1979

Symon, L. Flow thresholds in brain ischaemia and the effects of dr *Br J Anaesthesiol* , *57* 34, 1985

Tepperman, B L., and Jacobson, E D · Measurement of gastrointest blood flow *Annu Rev Physiol* , *44* 71, 1982

Wilkins, R H (ed) Cerebral Vasospasm Baltimore, Williams & kins, 1980

VI

THE BODY FLUIDS AND KIDNEYS

30 ■ Capillary Dynamics, and Exchange of Fluid Between the Blood and Interstitial Fluid

31 ■ The Lymphatic System, Interstitial Fluid Dynamics, Edema, and Pulmonary Fluid

32 ■ The Special Fluid Systems of the Body—Cerebrospinal, Ocular, Pleural, Pericardial, Peritoneal, and Synovial

33 ■ Partition of the Body Fluids: Osmotic Equilibria Between Extracellular and Intracellular Fluids

34 ■ Formation of Urine by the Kidney: Glomerular Filtration, Tubular Function, and Plasma Clearance

35 ■ Renal Mechanisms for (1) Controlling Glomerular Filtration, (2) Excreting Dilute or Concentrated Urine, and (3) Excreting Urea, Sodium, Potassium, and Fluid Volume

36 ■ Regulation of Blood Volume, Extracellular Fluid Volume, and Extracellular Fluid Composition by the Kidneys and by the Thirst Mechanism

37 ■ Regulation of Acid-Base Balance

38 ■ Renal Disease, Diuresis, and Micturition

30

Capillary Dynamics, and Exchange of Fluid Between the Blood and Interstitial Fluid

In tho capillaries the most purposeful function of the circulation occurs, namely, interchonge of nutrients and cellular excreta between the tissues and the circulating blood About 10 billion capillaries, which have a total surface area of probably 500 to 700 square meters, provide this function Indeed, it is rare that any single functional cell of the body is more than 20 to 30 microns away from a capillary.

The purpose of this chapter is to discuss the transfer of substances between the blood and interstitial fluid and especially to discuss the factors that affect the transfer of fluid volume itself between the circulating blood and the interstitial fluids

STRUCTURE OF THE CAPILLARY SYSTEM

The microcirculation of each organ is specifically organized to serve that organ's own special needs. However, in general, each nutrient artery entering an organ branches six to eight times before the arteries become small enough to be called "arterioles," which have diameters of generally less than 40 microns. Then the arterioles themselves branch two to five times, reaching diameters of about 9 microns at their ends where they supply blood to the capillaries

Figure 30–1 illustrates the structure of a representative capillary bed as seen in the mesentery, illustrating that blood enters the capillaries through an *arteriole* and leaves by way of a *venule* Blood from the arteriole passes into a series of *metarterioles*, which are called by some physiologists *terminal arterioles* and which have a structure midway between that of arterioles and capillaries After leaving the metarteriole, the blood enters the *capillaries*, some of which are large and are called *preferential channels* and others of which are small and are *true capillaries* After passing through the capillaries the blood enters the venule and returns to the general circulation

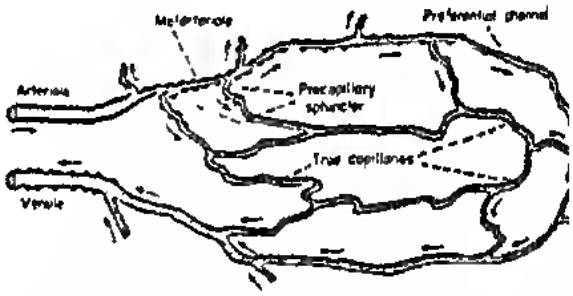


Figure 30–1. Structure of the mesenteric capillary bed (Zweifach Factors Regulating Blood Pressure New York, Macy, Jr., Foundation, 1950)

The arterioles are highly muscular, and their di ters can change manyfold, as was discussed in Ch 18 The metarterioles (the terminal arterioles) do have a continuous muscular coat, but smooth m fibers encircle the vessel at intermediate point illustrated in Figure 30–1 by the large black dots t sides of the metarteriole.

At the point where the true capillaries originate the metarterioles a smooth muscle fiber usually enc the capillary This is called the *precapillary sphin* This sphincter can open and close the entrance t capillary

The venules are considerably larger than the a ioles and have a much weaker muscular coat Y must be remembered that the pressure in the ver is much less than that in the arterioles so that venules can still contract considerably.

This typical arrangement of the capillary bed is found in all parts of the body; however, some sin arrangement is found, beginning with arterioles, l ing through metarterioles and capillaries, and retur through venules. Most important of all, the meta ioles (and the precapillary sphincters when they exist) are in extremely close contact with the tis that they serve. Therefore, the local conditions o tissues—the concentrations of nutrients, end produc

metabolism, hydrogen ions, and so forth—can cause direct effects on these in controlling the local blood flow in each minute tissue area.

Structure of the Capillary Wall. Figure 30–2 illustrates the ultramicroscopic structure of a typical capillary wall as found in most organs of the body, especially in the muscles and connective tissue. Note that the wall is composed of a unicellular layer of endothelial cells and is surrounded by a basement membrane on the outside. The total thickness of the wall is about 0.5 micron.

The diameter of the capillary is 4 to 9 microns, barely large enough for red blood cells and other blood cells to squeeze through.

"Pores" in the Capillary Membrane. Studying Figure 30–2, one sees two minute passageways connecting the interior of the capillary with the exterior. One of these is the *intercellular cleft*, which is a thin slit that lies between adjacent endothelial cells. The cells are held apart approximately 6 to 7 nanometers (60 to 70 Angstroms) by loose reticular fibrillae composed of proteoglycans, mainly hyaluronic acid. These intercellular clefts are located only at the edges of the endothelial cells and therefore represent no more than one thousandth of the total surface area of the capillary. Nevertheless, it is believed that most water-soluble ions and molecules pass between the interior and exterior of the capillary through these slit-pores.

Also present in the endothelial cells are many *pinocytic vesicles*. These form at one surface of the cell and move to the opposite surface where they discharge their contents, often carrying large molecules and even solid particles through the capillary membrane. Note also in Figure 30–2 that some of these pinocytic vesicles occasionally coalesce with each other and form a continuous channel through the endothelial membrane as illustrated by the so-called *pinocytic channel* shown to the right in the figure. However, it is still doubted whether large amounts of substances pass through the capillary membrane via such "pores" as this.

The "pores" in the capillaries of some organs have special characteristics to meet the peculiar needs of the organs. Some of these are:

(1) In the *brain, the junctions between the capillary endothelial cells are mainly "tight" junctions that will allow only very small molecules to pass into the brain tissues. This is called the blood-brain barrier* and will be discussed further in Chapter 32.

(2) In the *liver*, exactly the opposite is true: the clefts between the capillary endothelial cells are very wide open so that almost all dissolved substances of the plasma, including even the plasma proteins, can pass from the blood into the liver tissues. The pores of the intestinal membranes are midway between those of the muscles and the liver.

(3) In the *glomerular tufts of the kidney*, numerous small oval windows called *fenestrae* penetrate directly through the very middle of the endothelial cells, so that tremendous amounts of substances can be filtered through the glomeruli without having to pass through the clefts between the endothelial cells.

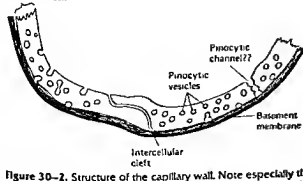


Figure 30–2. Structure of the capillary wall. Note especially the *intercellular cleft* at the junction between adjacent endothelial cells, it is believed that most water-soluble substances diffuse through the capillary membrane along this cleft

FLOW OF BLOOD IN THE CAPILLARIES—VASOMOTION

Blood usually does not flow continuously through the capillaries. Instead, it flows intermittently, turning on and off every few seconds or minutes. The cause of this intermittency is the phenomenon called *vasomotion*, which means intermittent contraction of the metarterioles and precapillary sphincters.

Regulation of Vasomotion. The most important factor found thus far to affect the degree of opening and closing of the metarterioles and precapillary sphincters is the concentration of oxygen in the tissues. When the oxygen concentration is very low, the intermittent periods of blood flow occur more often, and the duration of each period of flow lasts for a longer time, thereby allowing the blood to carry increased quantities of oxygen (as well as other nutrients) to the tissues. It follows also that the greater the use of oxygen by the tissue, the greater will be the amount of blood that flows. Thus, by this intermittent opening and closing of the precapillary sphincters and metarterioles, the blood flow to the tissue is *autoregulated*. This phenomenon, as well as multiple other factors that control local blood flow, was discussed in Chapter 20.

AVERAGE FUNCTION OF THE CAPILLARY SYSTEM

Despite the fact that blood flow through each capillary is intermittent, so many capillaries are present in the tissues that their overall function becomes averaged. That is, there is an *average rate of blood flow* through each tissue capillary bed, an *average capillary pressure* within the capillaries, an *average rate of transfer of substances* between the blood of the capillaries and the surrounding interstitial fluid. In the remainder of this chapter, we will be concerned with these averages, though one must remember that the average functions are in reality the functions of literally billions of in-

dividual capillaries, each operating intermittently in response to the local conditions in the tissues.

EXCHANGE OF NUTRIENTS AND OTHER SUBSTANCES BETWEEN THE BLOOD AND INTERSTITIAL FLUID

DIFFUSION THROUGH THE CAPILLARY MEMBRANE

By far the most important means by which substances are transferred between the plasma and interstitial fluids is by *diffusion* Figure 30–3 illustrates this process, showing that as the blood traverses the capillary, tremendous numbers of water molecules and dissolved particles diffuse back and forth through the capillary wall, providing continual mixing between the interstitial fluids and the plasma, as was explained in Chapter 9. Diffusion results from thermal motion of the water molecules and the dissolved substances in the fluid, the different particles moving first in one direction, then another, moving randomly in every direction.

Diffusion of Lipid-Soluble Substances Through the Capillary Membrane. If a substance is lipid-soluble, it can diffuse directly through the cell membranes of the capillary without having to go through the pores Such substances include especially oxygen and carbon dioxide. Since these can permeate all areas of the capillary membrane, their rates of transport through the capillary membrane are many times the rates for most lipid-insoluble substances such as sodium ions, glucose, and so forth.

Diffusion of Water Molecules Through the Capillary Membrane. After the lipid-soluble molecules, water molecules are the next most rapid to diffuse through the capillary membrane Much of the water diffusion occurs directly through t endothelial cell wall, first into the intracellul fluid of the endothelial cell and then out t membrane on the other side of the cell. In additio still much more water diffuses through the "pore in the capillary membrane, especially through t intercellular clefts

Diffusion of Water-Soluble, Lipid-Insolub Substances Through the Capillary Membran Many substances needed by the tissues are solub in water but cannot pass through the lipid mem branes of the endothelial cells; such substanc include sodium ions, chloride ions, glucose, and forth These substances diffuse between the plasm and interstitial fluids only through the capilla pores

Despite the fact that not over 1/1000 of t surface area of the capillaries is represented the intercellular junctions, the velocity of therm motion is so great that even this small area sufficient to allow tremendous diffusion of wat and water-soluble substances through these por To give one an idea of the extreme rapidity wi which substances diffuse, *the rate at which wat molecules diffuse through the capillary membra is approximately 80 times as great as the rate which plasma itself flows linearly along the cap lary* That is, the water of the plasma is exchang with the water of the interstitial fluids 80 tim before the plasma can go the entire distan through the capillary.

Effect of Molecular Size on Passa Through the Pores. The width of the capilla intercellular slit-pores, 6 to 7 nanometers, is abo 20 times the diameter of the water molecule, whi is the smallest molecule that normally pass through the capillary pores. On the other han the diameters of plasma protein molecules a slightly greater than the width of the pores. Oth substances, such as sodium ions, chloride ion glucose, and urea, have intermediate diamete Therefore, it is obvious that the permeability the capillary pores for different substances w vary according to their molecular diameters.

Table 30–1 gives the relative permeabilities the capillary pores in muscle for substances co monly encountered by the capillary membran

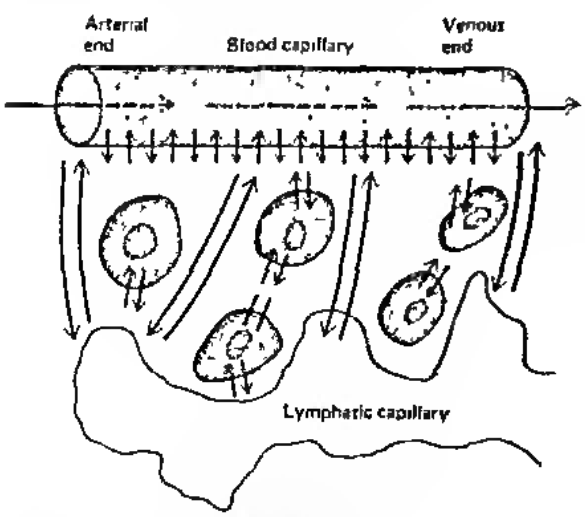


Figure 30–3. Diffusion of fluid and dissolved substances between the capillary and interstitial fluid spaces

Table 30–1 RELATIVE PERMEABILITY OF MUSCLE CAPILLARY PORES TO DIFFERENT-SIZED MOLECULES

Substance	*Molecular Weight*	*Permeabil*
Water	18	1 00
NaCl	58 5	0 96
Urea	60	0 8
Glucose	180	0 6
Sucrose	342	0 4
Inulin	5,000	0.2
Myoglobin	17,600	0 03
Hemoglobin	68 000	0 01
Albumin	69,000	<0 0001

Modified from Pappenheimer

illustrating, for instance, that the permeability for glucose molecules is 0.6 times that for water molecules, whereas the permeability for albumin molecules is less than 1/10,000 that for water molecules. Thus, the membrane is almost impermeable to albumin, which causes a significant concentration difference to develop between the albumin of the plasma and that of the interstitial fluid, as *will become evident later in the chapter.*

However, a word of caution must be issued at this point: The capillaries in different tissues have extreme differences in their permeabilities. For instance, the membrane of the liver capillary sinusoids is so permeable that even plasma proteins leak through the walls of these sinusoids almost as easily as water and other substances. Also, the permeability of the renal glomerular membrane for water and electrolytes is about 500 times the permeability of the muscle capillaries, but the glomerular and muscle permeabilities for protein are about the same. When we study these different organs later in this text, it will become clear why some tissues require greater degrees of capillary permeability than others—the liver, for instance, to transfer tremendous amounts of nutrients between the blood and the liver parenchymal cells, and the kidneys to allow filtration of large quantities of fluid for the formation of urine.

Effect of Concentration Difference on Net Rate of Diffusion Through the Capillary Membrane. As explained in Chapter 9, the "net" rate of diffusion of a substance through any membrane is proportional to the *concentration difference* between the two sides of the membrane. *That is,* the greater the difference between the concentrations of any given substance on the two sides of the capillary membrane, the greater will be the *net* movement of the substance through the membrane. Thus, the concentration of oxygen in the blood is normally greater than that in the interstitial fluids. Therefore, large quantities of oxygen normally move from the blood toward the tissues. Conversely, the concentration of carbon dioxide is greater in the tissues than in the blood, which obviously causes carbon dioxide to move into the blood and to be carried away from the tissues.

Fortunately, the rates of diffusion through the capillary membranes of most nutritionally important substances are so great that only slight concentration differences suffice to cause more than adequate transport between the plasma and interstitial fluid. For instance, the concentration of oxygen in the interstitial fluid immediately outside the capillary is probably no more than 1 per cent less than the concentration in the blood, and yet this 1 per cent difference causes enough oxygen to move from the blood into the interstitial spaces to provide all the oxygen required for tissue metabolism.

Effect of Hydrostatic Pressure Difference on Movement of Substances Through the Capillary Membrane—Bulk Flow

When the hydrostatic pressure is different on the two sides of the capillary membrane, far more water and dissolved substances move through the membrane than can be accounted for by increased diffusion alone. The reason this occurs is that the water and other substances *actually flow in bulk through the capillary pores when* a hydrostatic pressure difference occurs. This is the same phenomenon that occurs in a blood vessel when a pressure difference develops between the two ends of the vessel, namely, flow of the blood in *bulk* along the axis of the blood vessel from the high pressure area to the low pressure area. This means simply that the molecules of the fluid move together rather than individually by diffusion. So also do fluid and its dissolved substances flow through the capillary pores when a hydrostatic pressure difference develops across the capillary membrane. Therefore, this phenomenon is called *bulk flow* of fluid through the membrane.

Effect of Osmotic Pressure Difference on Movement of Substances Through the Capillary Membrane—Bulk Flow Caused by This Means Also

When a substance has a molecular weight too great for it to move through the slit-pores of the capillary membrane, it creates osmotic pressure at the point where the movement of the molecule is impeded. This pressure then acts exactly as does a hydrostatic pressure difference across the pore to cause bulk flow through the pore. The osmotic substance that normally is most important in causing bulk flow through the capillary membrane is the plasma proteins, as *will be explained* later in the chapter.

THE INTERSTITIUM AND THE INTERSTITIAL FLUID

Approximately one sixth of the body consists of spaces between cells, which collectively are called the *interstitium.* The fluid in these spaces is the *interstitial fluid.*

The structure of the interstitium is illustrated in Figure 30–4. It has two major types of solid structures: (1) collagen fiber bundles and (2) proteoglycan filaments. Figure 30–4 illustrates that the collagen fiber bundles extend long distances in the interstitium. They are extremely strong and therefore provide most of the tensional strength of the tissues. The proteoglycan filaments, on the other hand, are extremely thin, coiled molecules composed of about 98 per cent hyaluronic acid and 2 per cent protein. These molecules are so thin that they can never be seen with a light microscope and are very difficult to demonstrate even with the electron microscope. Nevertheless, they form a mat of very fine reticular filaments aptly described

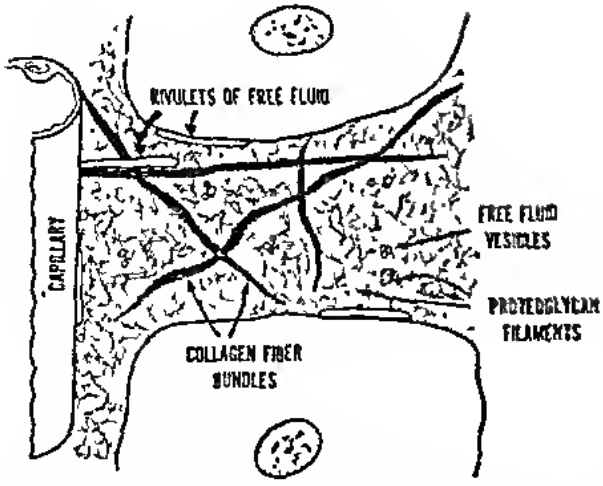


Figure 30–4. Structure of the interstitium. Proteoglycan filaments fill the spaces between the collagen fiber bundles. Free fluid vesicles are seen and small amounts of free fluid in the form of rivulets.

as a "brush pile." This brush pile of proteoglycan filaments is everywhere in the interstitium. It fills all the spaces between the collagen fibers, the crannies between the cells, and almost all the other minute spaces of the tissues.

"Tissue Gel" In the Interstitium. The fluid in the interstitium is an ultrafiltrate of plasma derived from the capillaries. In the interstitium this fluid is mainly entrapped in the minute spaces among the proteoglycan filaments. This combination of the proteoglycan filaments and the fluid entrapped within them has the characteristics of a *gel*, and therefore is often called the *tissue gel*.

Because of the large number of proteoglycan filaments, fluid *flows* through the tissue gel only very poorly. Instead, it mainly *diffuses* through the gel; that is, it moves molecule by molecule from one place to another by the process of kinetic motion rather than by large numbers of molecules moving together.

In contrast to the severe impediment of flow through the gel, diffusion through the gel occurs about 95 to 99 per cent as rapidly as it does through free fluid. This diffusion allows rapid transport through the interstitium not only of the water molecules, but also electrolytes, nutrients, cellular excreta, oxygen, carbon dioxide, and so forth.

"Free" Fluid in the Interstitium. Though almost all the fluid in the interstitium is normally entrapped within the tissue gel, occasionally small *rivulets of "free" fluid* and also small *free fluid vesicles* are also present, which means fluid that is free of the proteoglycan molecules and therefore can flow freely. When a dye is injected into the circulating blood, it can often be seen to flow [illegible] the interstitium in small rivulets, usually [illegible] along the surfaces of collagen fibers or [illegible]faces of cells. However, the amount of "free" [illegible] present in *normal* tissues is very slight. On the other hand, when the tissues develop *edema*, *it is the small pockets and rivulets of free fluid that expand tremendously.*

Relationship of the Free Fluid to the Gel Fluid. The free fluid and the gel fluid are continually interchanging with each other by diffusion. At the surface between the free fluid and the gel there is a slight osmotic pressure difference. Though this has never been measured accurately, the measurements that are available suggest that the osmotic pressure of the gel is about 2 mm Hg greater than that of the free fluid. This osmotic pressure is caused mainly by the Donnan equilibrium effect, which can be explained as follows: The proteoglycan molecules have a strong negative charge and therefore attract large numbers of positively charged ions, especially sodium ions, into the meshwork of filaments. These sodium ions are responsible for almost all the osmotic pressure of the gel.

Because of the uncertainty of the quantitative value of the osmotic pressure of the gel and therefore the uncertainty of the hydrostatic pressure of the fluid between the gel filaments, in most of the discussions of this chapter and the next we will speak mainly of the free fluid pressure rather than the intragel fluid pressure. There are other reasons for this as well: First, the fluid as it enters the interstitium from the capillary is in the form of free fluid. Second, as it leaves the interstitium to enter the lymphatic vessels the fluid is also in the form of free fluid. Third, when acute edema develops, the free fluid portion of the tissue mainly swells, and this is the most important clinical consideration relating to the interstitial fluids.

Nevertheless, as we quote pressures for the free fluid in the interstitium, the student should remember that the pressures in the minute spaces of the gel reticulum probably average about 2 mm Hg more positive than the pressures of the free fluid.

DISTRIBUTION OF FLUID VOLUME BETWEEN THE PLASMA AND INTERSTITIAL FLUID

The pressure in the capillaries continuously tends to force fluid and its dissolved substances through the capillary pores into the interstitial spaces. But, in contrast, osmotic pressure caused by the plasma proteins (called *colloid* osmotic pressure) tends to cause fluid movement by osmosis from the interstitial spaces into the blood; it is mainly this osmotic pressure that prevents continual loss of fluid volume from the blood into the interstitial spaces. Yet, the process is much more complicated than this and includes the role of the lymphatic system to return back to the circulation the small amounts of protein and fluid that do leak continuously into the interstitial spaces. In the following few paragraphs we will discuss all of the

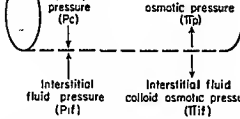


Figure 30–5. Forces operative at the capillary membrane tending to move fluid either outward or inward through the membrane.

factors that play a significant role in the movement of fluid volume through the capillary membrane, and in the following chapter we will discuss the role of the lymphatic system in this overall mechanism.

The Four Primary Forces That Determine Fluid Movement Through the Capillary Membrane. Figure 30–5 illustrates the four primary forces that determine whether fluid will move out of the blood into the interstitial fluid or in the opposite direction; those are:

1. The *capillary pressure* (Pc), which tends to move fluid outward through the capillary membrane.
2. The *interstitial fluid pressure* (Pif), which tends to move fluid inward through the capillary membrane when Pif is positive but outward when Pif is negative.
3. The *plasma colloid osmotic pressure* (πp), which tends to cause osmosis of fluid inward through the membrane.
4. The *interstitial fluid colloid osmotic pressure* (πif), which tends to cause osmosis of fluid outward through the membrane.

The regulation of fluid volumes in the blood and interstitial fluid is so important that each of these factors is discussed in turn in the following sections.

CAPILLARY PRESSURE

Unfortunately, the exact capillary pressure is not known because it has been impossible to measure capillary pressure under absolutely normal conditions. Yet two different methods have been used to estimate the capillary pressure: (1) *direct cannulation of the capillaries*, which has given an average mean capillary pressure of about 25 mm Hg, and (2) *indirect functional measurement of the capillary pressure*, which has given a capillary pressure averaging about 17 mm Hg. These methods are the following:

Cannulation Method for Measuring Capillary Pressure. To measure pressure in a capillary by cannulation, a microscopic glass pipet is thrust directly into the capillary, and the pressure is measured by an appropriate micromanometer system. Using this method, capillary pressures have been measured in capillaries of exposed tissues of lower animals and in large capillary loops of the eponychium at the base of the fingernail in human beings. These measurements have given pressures of 30 to 40 mm Hg in the arterial ends of the capillaries, 10 to 15 mm Hg in the venous ends, and about 25 mm Hg in the middle.

Isogravimetric and Isovolumetric Methods for Indirectly Measuring Mean Capillary Pressure. Figure 30–6 illustrates an *isogravimetric* method for estimating capillary pressure. This figure shows a section of gut held by one arm of a gravimetric balance. Blood is perfused through the gut. When the arterial pressure is decreased, the resulting decrease in capillary pressure allows the osmotic pressure of the plasma proteins to cause absorption of fluid out of the gut wall and makes the weight of the gut decrease. This immediately causes displacement of the balance arm. However, to prevent this weight decrease, the venous pressure is raised an amount sufficient to overcome the effect of decreasing the arterial pressure. In other words, the capillary pressure is kept constant by raising the venous pressure.

In the lower part of the figure, the changes in arterial and venous pressures that exactly nullify all weight changes are illustrated. The arterial and venous lines meet each other at a value of 17 mm Hg. Therefore, the capillary pressure must have remained at this same level of 17 mm Hg throughout those maneuvers, or otherwise filtration or absorption of fluid through the capillary walls would have occurred. Thus, in a roundabout way, the "functional" capillary pressure is measured to be about 17 mm Hg.

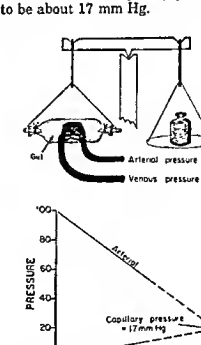


Figure 30–6. Isogravimetric method for measuring capillary pressure (explained in the text).

The *isovolumetric* method for measuring capillary pressure is essentially the same as the isogravimetric method, except that the *volume* of the tissue is recorded rather than the weight. The arterial and venous pressures are gradually brought toward each other and are continually adjusted so that the tissue neither gains nor loses volume. On extrapolating the arterial and venous pressures, one again finds that the capillary pressure estimated by this means is considerably lower than that measured by direct cannulation, again averaging about 17 mm Hg.

Functional Capillary Pressure. Since the cannulation method does not give the same pressure measurement as the isogravimetric and isovolumetric methods, one must decide which of these measurements is probably the true functional capillary pressure of the tissues. The isovolumetric and isogravimetric measurements are probably much nearer to the normal values for capillary pressure than are the micropipet measurements, for several reasons.

First, the metarterioles and precapillary sphincters of the capillary system are normally closed during a greater part of the vasomotion cycle than they are open. When they are closed the pressure in the entire capillary system beyond the closures should be almost exactly equal to the pressure at the venous ends of the capillaries, about 10 mm Hg. Therefore, if one considers a *weighted* average of the pressures in all capillaries, one would expect the *functional* mean capillary pressure to be much nearer to the pressure in the venous ends of the capillaries than to the pressure in the arterial ends.

Second, the surface area of the venous capillaries is several times as great as the surface area of the arterial capillaries. Therefore, the mean pressure in the venous capillaries, 10 mm Hg, plays far more role in determining the *functional* capillary pressure than does the mean arterial capillary pressure, 35 mm Hg

Third, the venous capillaries are several times as permeable as the arterial capillaries Therefore, for determining fluid movement through the capillary membrane, the venous capillary pressures are much more important than the arterial capillary pressures

Thus, there are many reasons for believing *that the normal functional mean capillary pressure is about 17 mm Hg*

What is the *average* functional pressure in the arterial ends of the capillaries? This we do not know, but it must be considerably below the 30 to 40 mm Hg measured in the micropipet in open arterial capillaries, because even the arterial ends [illegible] the capillaries are closed off most of the time [illegible] cause of metarteriole and precapillary sphincter [illegible] re We might estimate this pressure to be about 25 mm Hg, which is somewhat below the pipet measurements and somewhat above the functional mean capillary pressure.

INTERSTITIAL FLUID PRESSURE—INTRAGEL PRESSURE AND FREE FLUID PRESSURE

The interstitial fluid pressure, like capilla pressure, has been difficult to measure, priman because the maximum width of the spaces betwe the reticular fibers that make up the solid stru ture of the interstitium is only 10 to 40 nanom ters, much too small to cannulate for direct mea urement of the pressure. Therefore, mo frequently, indirect methods have been used f measuring this pressure.

Prior to 1961, it had been believed universal that the interstitial fluid pressure was alwa slightly greater than atmospheric pressure. Ho ever, on the basis of measurements made in o laboratory in 1961, we concluded that the interst tial fluid pressure is probably subatmospheric most tissues of the body, usually 2 to 7 mm less than the pressure of the surrounding air. Sin that time, a continuing controversy has exist concerning whether or not the interstitial fl pressure is truly less than the atmospheric leve

Also, the question must be asked: Is it t pressure of the fluid inside the gel, the *intra pressure,* that is important, or is it the pressure the free fluid, the *free fluid pressure,* that is i portant? From the point of view of understandi tissue function, the answer is that both of the are important. However, from the point of view the practice of medicine, it is the free fluid pressu that determines whether or not the tissues will nonedematous or edematous. Therefore, most tempts to measure the interstitial fluid pressu have been concerned primarily with measuri the free fluid pressure, and that is what we w be considering in the present chapter rather th the intragel pressure. As discussed earlier in t chapter, measurements of the osmotic pressure gel have indicated that the intragel fluid pressu is probably about 2 mm Hg more positive than t free fluid pressure.

With this as background, let us now summar some of the methods that have been used f measuring the interstitial free fluid pressure.

Measurement of Interstitial Free Fluid Pre sure in Implanted Perforated Hollow Ca aules. Figure 30–7 illustrates an indirect meth for measuring interstitial fluid pressure whi may be explained as follows· A small hollow plas capsule perforated by several hundred small hol is implanted in the tissues, and the surgical wou is allowed to heal for approximately one mont At the end of that time, tissue will have grov inward through the holes to line the inner surfa of the sphere. Furthermore, the cavity is fill with fluid that flows freely through the perfor tions back and forth between the fluid in t interstitial spaces and the fluid in the cavit Therefore, the pressure in the cavity should equ the free fluid pressure in the interstitial flu spaces. A needle is inserted through the skin a

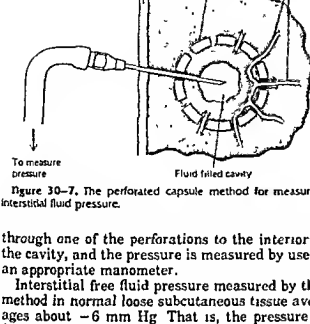


Figure 30–7. The perforated capsule method for measuring interstitial fluid pressure.

through one of the perforations to the interior of the cavity, and the pressure is measured by use of an appropriate manometer.

Interstitial free fluid pressure measured by this method in normal loose subcutaneous tissue averages about −6 mm Hg That is, the pressure is *less than atmospheric pressure* or, in other words, is a semivacuum or a suction.

Recently, instead of using capsules with large perforations as illustrated in Figure 30–6, we have made these capsules of a porous plastic material that has pora sizes too small for tissues to grow into the cavity of the capsule Using this new procedure, the pressures that have been measured have averaged −4 to −5 mm Hg.

The significance of the negativity of interstitial free fluid pressure is that it causes *suction of fluid out of the capillaries*, as we shall see in subsequent sections of this chapter. The mechanism for development of the negative interstitial fluid pressure is discussed in the following chapter in relation to interstitial fluid dynamics.

Measurement of Interstitial Free Fluid Pressure in Fluid Spaces Developed Underneath the Skin. Figure 30–8 illustrates another method that has been used to measure the interstitial fluid pressure. A small vacuum cup is placed over the skin, and the skin is sucked up into the cup and glued in the cup, then the vacuum is removed. Gradually, over about 24 hours, free fluid collects beneath the mound of skin, and its pressure comes to equilibrium. A needle is then inserted into the tissue so that its tip is within this free fluid,

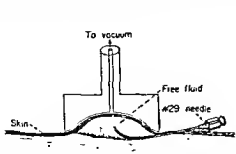


Figure 30–8. The equilibrium method for measuring interstitial fluid pressure.

and the pressure is recorded. Since the fluid is in free communication with the fluid in the surrounding tissues, the pressure measured here should equal the pressure in the other interstitial spaces. The pressure measured by this method in loose subcutaneous tissue is also negative, averaging between −2 and −8 mm Hg.

Measurement of Interstitial Free Fluid Pressure by Means of a Cotton Wick. A new method used recently is to insert into a tissue a small Teflon tube with about eight cotton fibers protruding from its end The cotton fibers form a "wick" that makes excellent contact with the tissue fluids and transmits interstitial fluid pressure into the Teflon tube; the pressure can then be measured from the tube by usual manometric means Pressures measured by this technique in loose subcutaneous tissue have also been negative, usually measuring about −2 mm Hg

Measurement of the Interstitial Fluid Pressure by Means of Tissue Cannulas. A method that was used for many years to "measure" interstitial fluid pressure was to insert a needle into the tissue, inject a small bolus of fluid through the tip of the needle, and then measure the pressure in the fluid bolus before it was absorbed The result was always a pressure measurement slightly greater than atmospheric pressure in subcutaneous tissue, usually between +1 and +4 mm Hg However, this was a false measurement of pressure because the preliminary injection of the bolus of fluid caused local tissue edema

Recently, microscopic pipets have been placed into tissue spaces in attempts to measure the interstitial fluid pressure. Earlier measurements in this way suggested slightly positive pressures of +1 to +2 mm Hg in the connective tissue of the bat wing, but negative pressures of −1 to −3 mm Hg have more recently also been made using this same method Since the pipets are 30 to 50 times as large as the spaces between the reticular fibrillae of the interstitium, one wonders how much the distortion of the tissues that is caused by the method affects the measurement

Interstitial Fluid Pressures in Tightly Encased Tissues

Some tissues of the body are surrounded by tight encasements, such as the cranial vault around the brain, the strong fibrous capsule around the kidney, the fibrous sheaths around the muscles, and the sclera around the eye In most of these, regardless of the method used for measurement, the interstitial fluid pressures are usually positive However, even so, these interstitial fluid pressures almost invariably are still less than the pressures exerted on the outsides of the tissues by their encasements For instance, the cerebrospinal fluid pressure surrounding the brain averages about +10 mm Hg, whereas the brain interstitial fluid pressure averages about +4 to +6 mm Hg. And in the kidneys the capsular pressure surrounding the kidney averages about +13 mm Hg whereas the reported interstitial fluid pressures have averaged about +6 mm Hg.

Thus, if one remembers that the pressure exerted on the skin is atmospheric pressure, which is considered to be zero pressure, then one might formulate a general rule that the normal interstitial fluid pressure is usually several millimeters of mercury negative with respect to the pressure that surrounds each tissue.

Is the True Interstitial Fluid Pressure in Subcutaneous Tissue Subatmospheric?

The concept that the interstitial fluid pressure is subatmospheric in many tissues of the body began with clinical observations that could not be explained by the previously held concept that interstitial fluid pressure was always positive. However, in addition to the clinical observations, many different special experiments in both animals and human beings have supported the concept that the interstitial fluid pressure is usually negative in both loose subcutaneous tissue and in most other soft tissues that are not surrounded by high pressure containments. Some of the pertinent observations are the following:

(1) When a skin graft is placed on a concave surface of the body, such as in an eye socket after removal of the eye, before the skin becomes attached to the sublying socket fluid tends to collect underneath the graft. Also, the skin attempts to shorten, which tries to pull it away from the concavity. Nevertheless, some negative force underneath the skin causes absorption of the fluid and usually will literally pull the skin back into the concavity.

(2) Less than 1 mm Hg of positive pressure is required to inject tremendous volumes of fluid into very loose subcutaneous tissues, such as beneath the lower eyelid, in the axillary space, and in the scrotum. Amounts of fluid calculated to be more than 100 times the amount of fluid normally in the interstitial space, when injected into these areas, will cause no more than about 2 mm Hg of positive pressure. The importance of these observations is that they show that such tissues do not have strong fibers that can prevent the accumulation of fluid. Therefore, some other mechanism must be available to prevent such fluid accumulation.

(3) In most natural cavities of the body where there is free fluid in dynamic equilibrium with the surrounding interstitial fluids, the pressures that have been measured have been negative. Some of these are the following:

Intrapleural space −8 mm Hg,
Joint synovial spaces −4 to −8 mm Hg, and
Epidural space −4 to −7 mm Hg

(4) The implanted capsule method for measuring the interstitial fluid pressure can be used to record dynamic changes in this pressure. The changes are approximately those that one would calculate to occur when (a) the arterial pressure is increased or decreased, (b) fluid is injected into the surrounding tissue spaces, or (c) a highly concentrated colloid osmotic agent is injected into the blood to absorb fluid from the tissue spaces. None of the other methods for measuring interstitial fluid pressure have shown this dynamic capability. Therefore, we tend to trust the capsule measurements much more than the other measurements of interstitial fluid pressure. The capsule measurements are usually more negative than the measurements made in other ways.

(5) The "cleaner" the implant of the capsules for measuring interstitial fluid pressure—that is, the less the fibrous reaction around the capsule—the more negative are the values recorded. Thus, the more normal the tissue, the more negative have been the measured values.

Therefore, in the remainder of this chapter and in the following two chapters, we will frequently speak of the concept of negative interstitial fluid pressures and will consider the pressure in the loose subcutaneous connective tissue to be an average value of −5.3 mm Hg, a value typical for the capsule measurements. Yet, the student should remember that some physiologists believe that all tissues have strong enough connective tissue fibers to maintain positive compression of the fluids in the interstitial spaces all the time.

PLASMA COLLOID OSMOTIC PRESSURE

Colloid Osmotic Pressure Caused by Proteins. The proteins are the only dissolved substances of the plasma and interstitial fluid that do not diffuse readily through the capillary membrane. Furthermore, when small quantities of protein do diffuse into the interstitial fluid, these are soon removed from the interstitial spaces by way of the lymph vessels. Therefore, the concentration of protein in the plasma averages about three times as much as that in the interstitial fluid, 7.3 gm/dl in the plasma versus 2 to 3 gm/dl in the interstitial fluid.

In the discussion of osmotic pressure in Chapter 4, it was pointed out that only those substances that fail to pass through the pores of a semipermeable membrane exert osmotic pressure. Since the proteins are the only dissolved constituents that do not readily penetrate the pores of the capillary membrane, it is the dissolved proteins of the plasma and interstitial fluids that are responsible for the osmotic pressure at the capillary membrane. To distinguish this osmotic pressure from that which occurs at the cell membrane, it is called either *colloid osmotic pressure* or *oncotic pressure*. The term "colloid" osmotic pressure is derived from the fact that a protein solution resembles a colloidal solution despite the fact that it is actually a true solution. (The osmotic pressure that results at the cell membrane is often called *total osmotic pressure* to distinguish it from the colloid osmotic pressure because essentially all dissolved substances of the body fluids exert osmotic pressure at the cell membrane.)

The Donnan Effect on the Colloid Osmotic Pressure. The so-called *Donnan effect* causes the colloid osmotic pressure of the plasma to be about 50 per cent greater than that caused by the proteins alone. This results from the fact that the proteins are negative ions, and to balance these negative ions a large number of

ositively charged ions (cations), mainly sodium ions, re held near to, but usually not bound to, the electro- egative charges of the proteins. These extra cations, herefore, increase the number of osmotically active ubstances wherever the proteins occur, and they in- rease the osmotic pressure. Even more important, how- ver (for mathematical reasons that cannot be explained ere), the Donnan equilibrium effect becomes progres- ively more significant the higher the concentration of roteins. This means, as illustrated in Figure 30–9, that he initial few grams of protein in each 100 ml of plasma r interstitial fluid has much less colloid osmotic effect han do the next few grams.

Normal Values for Plasma Colloid Osmotic Pressure. The colloid osmotic pressure of normal luman plasma averages approximately 28 mm Hg; 9 mm of this is caused by the dissolved protein, nd 9 mm by the cations held in the plasma by he Donnan effect of the proteins, as was just liscussed.

Note particularly that the 28 mm Hg colloid osmotic pressure that can develop at the capillary nembrane is only 1/200 the total osmotic pressure hat would develop at a *cell* membrane if normal nterstitial fluid were on one side of the cell mem- brane and pure water on the other side. Thus, the colloid osmotic pressure of the plasma is actually a weak osmotic force; but, even so, it plays an exceedingly important role in the maintenance of normal blood and interstitial fluid volumes.

Effect of the Different Plasma Proteins on Colloid Osmotic Pressure. The plasma proteins are a mixture of proteins that contains albumin, with an average molecular weight of 69,000, globulins, 140,000; and fibrinogen, 400,000. Thus, 1 gram of globulin contains only half as many molecules as 1 gram of albumin, and 1 gram of fibrinogen contains only one sixth as many molecules as 1 gram of albumin (It will be recalled from the discussion of osmotic pressure in Chapter 4 that the osmotic pressure is determined by the *number of molecules* dissolved in a fluid rather than by the weight of these molecules.) The average relative concentrations of the different types of proteins in the plasma and their respective colloid osmotic pressures are·

	gm %	Π*p (mm Hg)*
Albumin	4.5	21.8
Globulins	2.5	6 0
Fibrinogen	0.3	0 2
TOTAL	7.3	28 0

Thus, about 75 per cent of the total colloid osmotic pressure of the plasma results from the albumin fraction, 25 per cent from the globulins and almost none from the fibrinogen. Therefore, from the point of view of capillary dynamics, it is mainly albumin that is important.

Figure 30–9 illustrates graphically the colloid osmotic pressures exerted by different concentrations of albumin and four different fractions of globulins.

INTERSTITIAL FLUID COLLOID OSMOTIC PRESSURE

Though the size of the usual capillary pore is smaller than the molecular sizes of the plasma

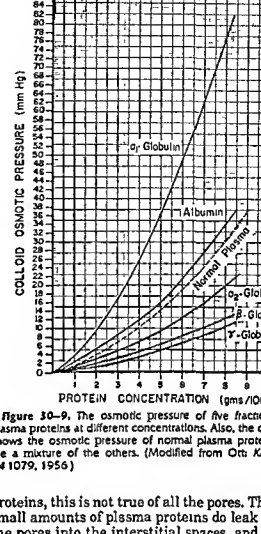


Figure 30–9. The osmotic pressure of five fractions of the plasma proteins at different concentrations. Also, the dashed line shows the osmotic pressure of normal plasma proteins, which are a mixture of the others. (Modified from Ott *Klin. Wschr.* *34* 1079, 1956)

proteins, this is not true of all the pores. Therefore, small amounts of plasma proteins do leak through the pores into the interstitial spaces, and perhaps additional amounts are transported by pinocytosis.

The total quantity of protein in the entire 12 liters of interstitial fluid of the body is slightly greater than the total quantity of protein in the plasma itself, but since this volume is four times the volume of plasma the average protein *concentration* of the interstitial fluid is only a little more than one fourth that in plasma, or approximately 2 grams per cent. Referring to the diagram in Figure 30–9 one finds that the average colloid osmotic pressure for this concentration of proteins in the interstitial fluids is approximately 6 mm Hg.

EXCHANGE OF FLUID VOLUME THROUGH THE CAPILLARY MEMBRANE

Now that the different factors affecting capillary membrane dynamics have been discussed, it is possible to put all these together to see how normal capillaries function.

The capillary pressure at the arterial ends of the capillaries is 15 to 25 mm Hg greater than at the venous ends. Because of this difference, fluid

"filters" out of the capillaries at their arterial ends and then is reabsorbed into the capillaries at their venous ends. Thus, a small amount of fluid actually "flows" through the tissues from the arterial ends of the capillaries to the venous ends. The dynamics of this flow are the following:

Analysis of the Forces Causing Filtration at the Arterial End of the Capillary. The forces operative at the arterial end of the capillary that cause movement through the capillary membrane are:

	mm Hg
Forces tending to move fluid outward:	
Capillary pressure	30.0
Negative interstitial free fluid pressure	5 3
Interstitial fluid colloid osmotic pressure	6 0
TOTAL OUTWARD FORCE	41 3
Force tending to move fluid inward:	
Plasma colloid osmotic pressure	28 0
TOTAL INWARD FORCE	28 0
Summation of forces:	
Outward	41.3
Inward	28 0
NET OUTWARD FORCE	13.3

Thus, the summation of forces at the arterial end of the capillary shows a net *filtration pressure* of 13.3 mm Hg, tending to move fluid in the outward direction.

This 13.3 mm Hg filtration pressure causes, on the average, about 0 5 per cent of the plasma to filter out of the arterial end of the capillaries into the interstitial spaces.

Analysis of Reabsorption at the Venous End of the Capillary. The low pressure at the venous end of the capillary changes the balance of forces in favor of absorption as follows:

	mm Hg
Force tending to move fluid inward:	
Plasma colloid osmotic pressure	28.0
TOTAL INWARD FORCE	28 0
Forces tending to move fluid outward:	
Capillary pressure	10 0
Negative interstitial free fluid pressure	5.3
Interstitial fluid colloid osmotic pressure	6 0
TOTAL OUTWARD FORCE	21.3
Summation of forces:	
Inward	28.0
Outward	21.3
NET INWARD FORCE	6.7

Thus, the force that causes fluid to move into the capillary, 28 mm Hg, is greater than that opposing reabsorption, 21.3 mm Hg. The difference, 6.7 mm Hg, is the *reabsorption pressure*. This reabsorption pressure is considerably less than the filtration pressure, but remember that the venous capillaries are more numerous and more permeable than the arterial capillaries so that less pressure is required to cause the inward movement of fluid.

The reabsorption pressure causes about nine tenths of the fluid that has filtered out of the arterial ends of the capillaries to be reabsorbed at the venous ends The other one tenth flows into the lymph vessels, as is discussed in the following chapter.

Flow of Fluid Through the Interstitial Spaces

The 0 5 per cent of the plasma fluid that filters out of the arterial ends of the capillaries *flows* through the tissue spaces to the venous ends of the capillaries where all but about one tenth of it is reabsorbed (A much higher proportion than this is reabsorbed in the muscles where very little protein leaks through the capillary membranes, and much less is reabsorbed in the liver where tremendous amounts of protein leak.)

Distinction Between Filtration and Diffusion. It is especially important to distinguish between *filtration* and *diffusion* through the capillary membrane Diffusion occurs in both directions, whereas filtration is the net movement of fluid out of the capillaries at the arterial ends. The rate of diffusion of water through all the capillary membranes of the entire body is about 240,000 ml/minute, whereas the normal rate of filtration at the arterial ends of all the capillaries is only 16 ml/minute, a difference of about 15,000-fold. Thus, the quantitative rate of diffusion of water and nutrients back and forth between the capillaries and the interstitial spaces is tremendous in comparison with the minute rate of "flow" of fluid through the tissues.

THE STARLING EQUILIBRIUM FOR CAPILLARY EXCHANGE

E. H. Starling pointed out almost a century ago that under normal conditions a state of near-equilibrium exists at the capillary membrane whereby the amount of fluid filtering outward through the arterial capillaries equals that quantity of fluid that is returned to the circulation by reabsorption at the venous ends of the capillaries This near-equilibrium is caused by near-equilibration of the *mean* forces tending to move fluid through the capillary membranes. If we assume that the mean *functional* capillary pressure is 17 mm Hg, the normal mean dynamics of the capillary are the following:

	mm Hg
Mean forces tending to move fluid outward:	
Mean capillary pressure	17.0
Negative interstitial free fluid pressure	5.3
Interstitial fluid colloid osmotic pressure	6.0
TOTAL OUTWARD FORCE	28.3
Mean force tending to move fluid inward:	
Plasma colloid osmotic pressure	28.0
TOTAL INWARD FORCE	28.0
Summation of mean forces:	
Outward	28.3
Inward	28.0
NET OUTWARD FORCE	0.3

Thus, we find a near-equilibrium but nevertheless a slight imbalance of forces, 0.3 mm Hg, that causes slightly more filtration of fluid into the interstitial spaces than reabsorption. This slight excess of filtration is called the *net filtration*, and it is balanced by fluid return to the circulation through the lymphatics. The normal rate of net filtration in the entire body is about 2 ml/min. This figure also represents the average rate of fluid flow into the lymphatics each minute.

The Filtration Coefficient. In the above example an average net imbalance of forces at the capillary membranes of 0.3 mm Hg causes a net rate of fluid filtration in the entire body of 2 ml/minute. Expressing this per millimeter of mercury, one finds a net filtration rate of 6.67 ml fluid/min/mm Hg for the entire body. This expression is the *filtration coefficient*.

The filtration coefficient can also be expressed for different areas of the body in terms of the rate of filtration/min/mm Hg/100 grams of tissue. On this basis the filtration coefficient of the average tissue is about 0.01 ml/min/mm Hg/100 grams of tissue. However, because of extreme differences in permeabilities of the capillary systems in different tissues, this coefficient varies more than 100-fold among the different tissues. It is very small in both brain and muscle, moderately great in subcutaneous tissue, large in the intestine, and extreme in the liver and the glomerulus of the kidney where the *pores are either numerous or wide open*. By the same token, the permeation of proteins through the capillary membranes varies greatly as well. The concentration of protein in the interstitial fluid of muscles is about 1.5 grams per cent, in subcutaneous tissue 2 grams per cent, in intestine 4 grams per cent, and in liver 6 grams per cent.

Effect of Abnormal Imbalance of Forces at the Capillary Membrane

If the mean capillary pressure rises above 17 mm Hg the net force tending to cause filtration of fluid into the tissue spaces obviously rises. Thus a 20 mm Hg rise in mean capillary pressure causes an increase in the net filtration pressure from 0.3 mm Hg to 20.3 mm Hg, which results in 68 times as much net filtration of fluid into the interstitial spaces as normally occurs, and this would require also 68 times the normal flow of fluid into the lymphatic system, an amount that is usually too much for the lymphatics to carry away. As a result, fluid begins to accumulate in the interstitial spaces, and edema results.

Conversely, if the capillary pressure falls very low, net reabsorption of fluid into the capillaries occurs instead of net filtration, and the blood volume increases at the expense of the interstitial fluid volume. The effects of these imbalances at the capillary membrane are discussed in the following chapter in relation to the formation of edema.

REFERENCES

Allen, P C, *et al* Plasma proteins. Analytical and preparative techniques. Philadelphia, J B Lippincott Co., 1977

Arendshorst, W J, and Gottschalk, C W Glomerular ultrafiltration dynamics. Historical perspective *Am J Physiol*, *248* F163, 1985

Aschheim, E Passage of substances across the walls of blood vessels Kinetics and mechanism *In* Kaley, G, and Altura, B M. (eds) Microcirculation. Vol 1 Baltimore, University Park Press, 1977, p 213

Baez, S Microcirculation *Annu Rev Physiol*, *39* 391, 1977

Barry, P H, and Diamond, J M Effects of unstirred layers on membrane phenomena *Physiol Rev*, *64* 763, 1984

Bang, D H (ed.) The Chemistry and Physiology of the Human Plasma Proteins, New York, Pergamon Press, 1979

Brace, R A, and Guyton, A C Effect of hindlimb isolation procedure on isogravimetric capillary pressure and transcapillary fluid dynamics in dogs. *Circ. Res*, *38*.192, 1976.

Brace, R A, and Guyton, A C Interaction of transcapillary Starling forces in the isolated dog forelimb *Am J Physiol*, *233* H136, 1977

Brace, R A, *et al*. Determinants of isogravimetric capillary pressure in the isolated dog hindlimb *Am J Physiol*, *233* H130, 1977

Bundgaard, M: Transport pathways in capillaries—in search of pores. *Annu. Rev Physiol*, *42* 325, 1980

Cameron, B, *et al*. Production and characterization of monoclonal antibodies directed against connective tissue proteoglycans *Fed. Proc*, *44*.386, 1985.

Comper, W. D, and Laurent, T C Physiological function of connective tissue polysaccharides. *Physiol. Rev*, *58* 255, 1978

Crone, C, and Christensen, O.: Transcapillary transport of small solutes and water. *In* Guyton, A. C, and Young, *D B (eds) International* Review of Physiology· Cardiovascular Physiology III. Vol 18 Baltimore, University Park Press, 1979, p. 149.

Crone, C., and Levitt, D G: Capillary permeability to small solutes. *In* Renkin, E. M, and Michel, C. C. (eds). Handbook of Physiology Sec 2, Vol IV Bethesda, American Physiological Society, 1984, p 411

Duling, B. R., and Berne, R. M· Propagated vasodilation in the microcirculation of the hamster cheek pouch *Circ Res*, *26* 163, 1970

Gabbiani, G, and Majno, G.. Fine structure and endothelium. *In* Kaley, G., and Altura, B M. (eds) Microcirculation. Vol 1 Baltimore, University Park Press, 1977, p 133

Gore, R. W, and McDonagh, P. F· Fluid exchange across single capillaries. *Annu. Rev Physiol*, *42* 337, 1980

Gross, J F, and Popel, A. (eds.). Mathematics of Microcirculation Phenomena. New York, Raven Press, 1980

Guyton, A. C.: Concept of negative interstitial pressure based on pressures in implanted perforated capsules *Circ Res*, *12* 399, 1963

Guyton, A. C Interstitial fluid pressure II Pressure volume curves of interstitial space *Circ Res*, *16* 452, 1965

Guyton, A C Interstitial fluid pressure-volume relationships and their regulation *In* Wolstenholme, G E W, and Knight, J (eds) Ciba Foundation Symposium on Circulatory and Respiratory Mass Transport. London, J & A Churchill Ltd, 1969 p. 4

Guyton, A C, *et al*. Interstitial fluid pressure III Its effect on resistance to tissue fluid mobility *Circ Res*, *19* 412, 1966

Guyton, A C, et al Interstitial fluid pressure IV Its effect on fluid movement through the capillary wall *Circ. Res*, *19* 1022, 1966

Guyton, A. C *et al.* Interstitial fluid pressure *Physiol Rev*, *51*.527, 1971

Guyton A C, *et al* Circulatory Physiology II Dynamics and Control of the Body Fluids Philadelphia, W B Saunders Co, 1975

Haddy, F J, Scott, J B, and Grega, G J Peripheral circulation. Fluid transfer across the microvascular membrane *Int Rev Physiol*, *9* 63, 1976

Hruza, Z. Connective tissue. *In* Kaley, G., and Altura B M (eds) Microcirculation. Vol 1 Baltimore, University Park Press, 1977, p 167

Intaglietta, M The measurement of pressure and flow in the microcirculation *Microvasc. Res*, *5* 357, 1973

Intaglietta, M Transcapillary exchange of fluid in single microvessels *In* Kaley, G, and Altura B M (eds) Microcirculation Vol 1 Baltimore, University Park Press, 1977, p 197

Johnson, P C The microcirculation and local and humoral control of the circulation *In* MTP International Review of Science Physiology Vol 1 Baltimore University Park Press, 1974, p 163

Landis, E M Capillary pressure and capillary permeability *Physiol Rev*, *14* 404, 1934

Landis E M, and Pappenheimer, J R Exchange of substances through the capillary walls *In* Hamilton, W F (ed) Handbook of Physiology Sec 2 Vol 2 Baltimore, Williams & Wilkins, 1963, p 961

Leonard, E F, and Jorgensen, S B The analysis of convection and diffusion in capillary beds *Annu. Rev Biophys Bioeng* 3 293, 1974

Meyerson H S The physiologic importance of lymph *In* Hamilton, W F (ed.) Handbook of Physiology Sec 2 Vol 2 Baltimore Williams & Wilkins, 1963, p 1035

Michel, C C Fluid movements through capillary walls *In* Renkin E M, and Michel, C C (eds) Handbook of Physiology Sec 2, Vol IV Bethesda, American Physiological Society 1984, p 375

Nicoll, P A, and Taylor, A E Lymph formation and flow *Annu Rev Physiol*, *39* 73, 1977

Pappenheimer, J R. Passage of molecules through capillary walls. *Physiol. Rev*, *33* 387, 1953

Pappenheimer, J R. Contributions to microvascular research of Jean Leonard Marie Poiseuille *In* Renkin, E M, and Michel, C C (eds) Handbook of Physiology Sec 2, Vol. IV Bethesda, American Physiological Society, 1984, p. 1

Peeters, H, and Wright, P H (eds). Plasma Protein Pathology A Workshop on Plasma Proteins Their Availability, Assay, and Therapeutic Uses New York, Pergamon Press, 1979

Prather, J W, *et al* Direct continuous recording of plasma colloid osmotic pressure of whole blood *J Appl Physiol*, *24* 602, 1968.

Prockop, D J Collagen, elastin, and proteoglycans: matrix for fluid accumulation in the lung *In* Fishman, A P., and Renkin, E M (eds) Pulmonary Edema. Baltimore, Waverly Press, 1979, p 125

Rhodin, J A G Architecture of the vessel wall *In* Bohr, D F, *et al* (eds) Handbook of Physiology Sec 2 Vol II Baltimore, Williams & Wilkins, 1980 p 1

Rothschild, M A, *et al* Albumin synthesis *In* Javitt, N B (ed) International Review of Physiology Liver and Biliary Tract Physiology I Vol 21 Baltimore, University Park Press, 1980, p 249

Sharon N Glycoproteins *Sci Am*, *230*(5) 78, 1974

Shepro D, and D'Amore, P A. Physiology and biochemistry of the vascular wall endothelium *In* Renkin, E M, and Michel, C C (eds) Handbook of Physiology Sec 2, Vol IV Bethesda, American Physiological Society, 1984, p 103

Simionescu, M Transendothelial movement of large molecules in the microvasculature *In* Fishman, A. P, and Renkin, E. M (eds) Pulmonary Edema Baltimore, Waverly Press, 1979, p 39

Simionescu, M, and Simionescu, N Ultrastructure of the microvascular wall Functional correlations *In* Renkin, E. M, and Michel, C. C. (eds) Handbook of Physiology Sec 2, Vol IV Bethesda, American Physiological Society, 1984, p 41

Simionescu, N Cellular aspects of transcapillary exchange *Physiol Rev*, *63* 1536, 1983.

Stromme, S B, *et al* Interstitial fluid pressure in terrestrial and semiterrestrial animals *J Appl Physiol*, *27* 123, 1969

Taylor, A E, and Granger, D N Exchange of macromolecules across the microcirculation *In* Renkin, E M., and Michel, C. C (eds) Handbook of Physiology Sec. 2, Vol. IV Bethesda, American Physiological Society, 1984, p 467

Wiedeman, M P Architecture *In* Renkin, E M, and Michel, C. C. (eds) Handbook of Physiology Sec 2, Vol IV, Bethesda, American Physiological Society, 1984, p 11

31

The Lymphatic System, Interstitial Fluid Dynamics, Edema, and Pulmonary Fluid

THE LYMPHATIC SYSTEM

The lymphatic system represents an accessory route by which fluids can flow from the interstitial spaces into the blood. And, most important of all, the lymphatics can carry proteins and large particulate matter away from the tissue spaces, neither of which can be removed by absorption directly into the blood capillary. We shall see that this removal of proteins from the interstitial spaces is an absolutely essential function, without which we would die within about 24 hours.

THE LYMPH CHANNELS OF THE BODY

With the exception of a very few, tissues of the body have lymphatic channels that drain excess fluid directly from the interstitial spaces The exceptions include the superficial portions of the skin, the central nervous system, deeper portions of peripheral nerves, the endomysium of muscles, and the bones. However, even these tissues have minute interstitial channels called *prelymphatics* through which interstitial fluid can flow; eventually this fluid flows into lymphatic vessels or, in the case of the brain, flows into the cerebrospinal fluid and thence directly back into the blood.

Essentially all the lymph from the lower part of the body—even most of that from the legs—flows up the *thoracic duct* and empties into the venous system at the juncture of the *left* internal jugular vein and subclavian vein, as illustrated in Figure 31–1. However, small amounts of lymph from the lower part of the body can enter the veins in the inguinal region and perhaps also at various points in the abdomen.

Lymph from the left side of the head, the left arm, and parts of the chest region also enters the thoracic duct before it empties into the veins. Lymph from the right side of the neck and head, from the right arm, and from parts of the thorax enters the *right lymph duct*, which then empties into the venous system at the juncture of the *right* subclavian vein and internal jugular vein.

The Lymphatic Capillaries and Their Permeability. Most of the fluid filtering from the arterial capillaries flows among the cells and is finally reabsorbed back into the *venous ends* of the *blood capillaries;* but about *one tenth* of the fluid enters the *lymphatic capillaries* instead and returns to the blood through the lymphatic system rather than through the venous capillaries.

The minute quantity of fluid that returns to the circulation by way of the lymphatics is extremely important because substances of high molecular weight such as proteins cannot be reabsorbed into the venous capillaries, but they can enter the lymphatic capillaries almost completely unimpeded. The reason for this is a special structure of the lymphatic capillaries, illustrated in Figure 31–2. This figure shows the endothelial cells of the capillary attached by *anchoring filaments* to the connective tissue between the surrounding tissue cells. At the junctions of adjacent endothelial cells there are usually very loose connections between the cells. Instead, the edge of one endothelial cell usually overlaps the edge of the adjacent one in such a way that the overlapping edge is free to flap inward, thus forming a minute valve that opens to the interior of the capillary. Interstitial fluid, along with its suspended particles, can push the valve open and flow directly into the capillary. But this fluid cannot leave the capillary once it has entered because any backflow will close the flap valve. Thus, the lymphatics have valves at the very tips of the terminal lymphatic capillaries as well as valves along their larger vessels up to the point where they empty into the blood circulation.

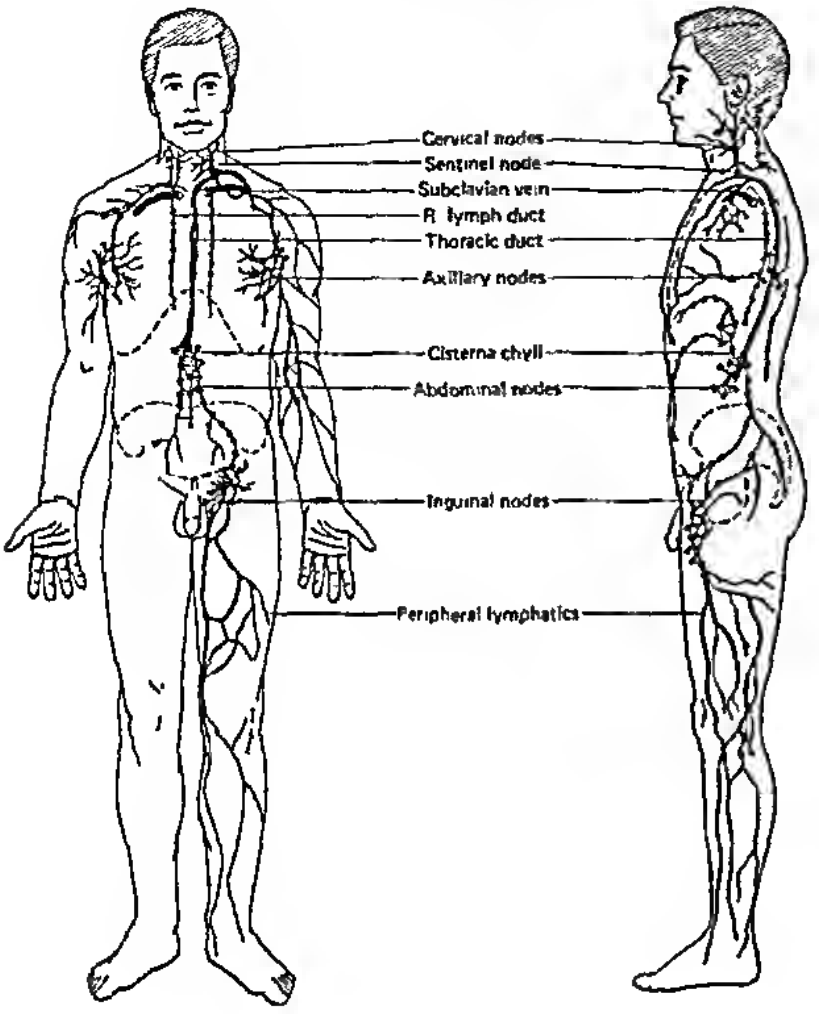


Figure 31–1. The lymphatic system.

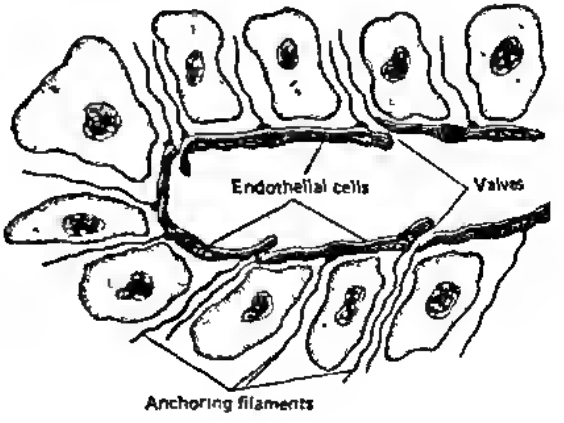


Figure 31–2. Special structure of the lymphatic capillaries that permits passage of substances of high molecular weight back into the circulation.

FORMATION OF LYMPH

Lymph is derived from interstitial fluid that flows into the lymphatics. Therefore, lymph as it first flows from each tissue has almost the same composition as the interstitial fluid in the part of the body from which the lymph flows.

The protein concentration in the interstitial fluid averages about 2 grams per cent, and the protein concentration of lymph flowing from most peripheral tissues is near this value. On the other hand, lymph formed in the liver has a protein concentration as high as 6 grams per cent, and lymph formed in the intestines has a protein concentration as high as 3 to 4 grams per cent. Since about two thirds of all lymph is derived from the liver and intestines, the thoracic lymph, which is a mixture of lymph from all areas of the body, usually has a protein concentration of 3 to 5 grams per cent.

The lymphatic system is also one of the major routes for absorption of nutrients from the gastrointestinal tract, being responsible principally for the absorption of fats, as will be discussed in Chapter 65. Indeed, after a fatty meal, thoracic duct lymph sometimea contains as much as 1 to 2 per cent fat.

Finally, even large particles, such as bacteria, can push their way between the endothelial cells of the lymphatic capillaries and in this way enter tha lymph. As tha lymph passes through the lymph nodes, these particles are removed and destroyed, as was discussed in Chapter 6.

The Mechanism hy Which Proteins Become Concentrated in Interatitial Fluids. Fluid filtering from tha arterial ends of the capillaries in peripheral tissues such as tha subcutaneous tissue usually has a protein concentration of about 0.2 per cent, though the averaga concentration of protein in tha interstitial fluids is almost ten times this value. The reason for this difference is that only a small proportion of the protein that leaks into the tissue spaces is reabsorbed at the venous ends of the capillaries, even though most of the filtered water and ions are reabsorbed In this way, the proteins become concentrated in the interstitial fluid before it flows into the lymphatics.

Additional Concentration of the Lymph in Lymph Nodes. As lymph passes through the lymph node, it is exposed to an extensive network of blood capillaries where large quantities of water and small molecular weight solutes, but not protein, diffuse freely back and forth between the blood and the nodal lymph. If the balance between the hydrostatic and colloid osmotic forces across the capillary membranes is not in equilibrium, bulk flow of water and small solutes will occur through the membrane in an attempt to bring about equilibrium. Usually almost complete equilibrium is established *before* the lymph leaves the efferent side of the node. For this reason, the protein concentration in the *efferent lymph* (lymph leaving the node) may be quite different from its concentration in the *offerent lymph* (lymph entering the node). The equilibrium concentration of protein in the efferent lymph is usually about 4 grams per cent, which is greater than that in the afferent lymph from most areas of the body. Therefore, most lymph entering the lymph nodes is concentrated as it passes through the nodes. On the other hand, afferent lymph from the liver, which has a very high concentration of protein, is normally diluted as it passes through a node.

Because many research workers collect lymph and use this as a measure of tissue fluid composition—and especially because most of the lymph is collected after it has passed through lymph nodes—it is important to recognize the difference between afferent and efferent lymph.

TOTAL RATE OF LYMPH FLOW

Approximately 100 ml of lymph flow through the thoracic duct of a resting man per hour, and perhaps another 20 ml of lymph flow into the circulation each hour through other channels, making a total estimated lymph flow of perhaps 120 ml per hour. This is about 1/120,000 of the calculated rate of fluid *diffusion* back and forth through the capillary membranes, and it is one tenth the rate of fluid *filtration* from the arterial ends of the capillaries into the tissua spaces in the entire body. Thesa facts illustrate that the flow of lymph is relatively small in comparison with the total exchange of fluid between the plasma and the interstitial fluid.

Factors That Determine the Rate of Lymph Flow

Interstitial Fluid Pressure. Elevation of interstitial *free* fluid pressure above its normal level of −5.3 mm Hg increases the flow of interstitial fluid into the lymphatic capillaries and consequently also increases the rate of lymph flow. The increase in flow becomes progressively greater as the interstitial fluid pressure rises until this pressure reaches a value slightly greater than 0 mm Hg; at that point the flow rate reaches a maximum, but by then it has risen to 10 to 50 times normal. Therefore, any factor (besides obstruction of the lymphatic system itself) that tends to increase lymphatic pressure increases the rate of lymph flow. Such factors include

Elevated capillary pressure,
Decreased plasma colloid osmotic pressure,
Increased interstitial fluid protein, and
Increased permeability of the capillaries.

The Lymphatic Pump. Valves exist in all lymph channels; typical valves are illustrated in Figure 31–3 in *collecting lymphatics* into which the lymphatic capillaries empty. In the large lymphatics, valves exist every few millimeters, and in

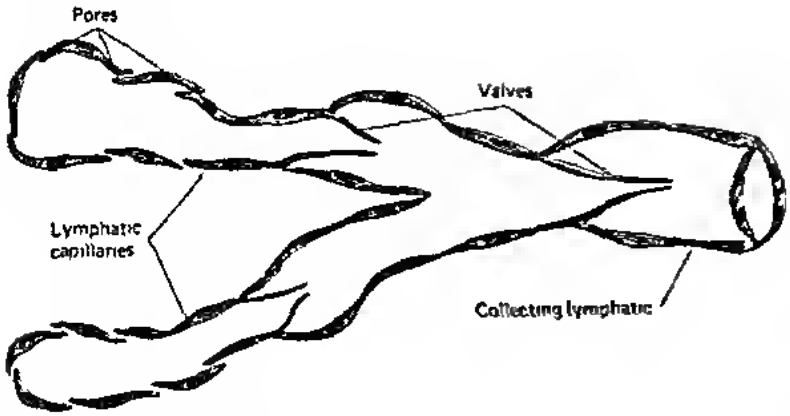


Figure 31–3. Structure of lymphatic capillaries and a collecting lymphatic, showing also the lymphatic valves.

the smaller lymphatics the valves are much closer together than this.

Intrinsic Pumping by the Lymph Vessel. Motion pictures of exposed lymph vessels, both in animals and human beings, show that when a lymph vessel becomes stretched with fluid the smooth muscle in the wall of the vessel automatically contracts. Furthermore, each segment of the lymph vessel between successive valves functions as a separate automatic pump. That is, filling of a segment causes it to contract, and the fluid is pumped through tha downstream valve into the following lymphatic segment. This fills the subsequent segment and a few seconds later it too contracts, the process continuing all along the lymphatic until the fluid is finally emptied In a large lymph vessel this lymphatic pump can generate pressures as high as 25 to 50 mm Hg if the outflow from the vessel becomes blocked.

Pumping Caused by External Compression of the Lymphatics. In addition to the pumping caused by intrinsic contraction of the lymph vessel walls, any external factor that compresses the lymph vessel can also cause pumping. In order of their importance, such factors are

Contraction of muscles,
Movements of the parts of the body,
Arterial pulsations, and
Compression of the tissues by objects outside the body.

Obviously, the lymphatic pump becomes very active during exercise, often increasing lymph flow as much as 10- to 30-fold On the other hand, during periods of rest lymph flow is very sluggish.

The Lymphatic Capillary Pump. Many physiologists believe that the lymphatic capillary is also capable of pumping lymph, in addition to the pump of the larger lymph vessels. As explained earlier in the chapter, the walls of lymphatic capillaries are tightly adherent to surrounding tissue cells by means of their anchoring filaments. Therefore, each time excess fluid enters the tissue and causes the tissue to swell, the anchoring filaments pull the lymphatic capillary open, and fluid flows into the capillary through the junctions between the endothelial cells. Then, when the tissue is compressed, the pressure inside the capillary increases and tends to push the fluid in two directions: first, backward through the openings between the endothelial cells, and, second, forward into the *collecting lymphatic*. However, since the edges of the endothelial cells normally overlap each other on the inside of the lymphatic capillary as illustrated in Figure 31–2, any backward flow through this opening closes the overlapping cell over the opening. Thus the opening closes as if it were a one-way valve and very little fluid flows backward into tha tissues. On the other hand, the lymph that travels forward into the collecting lymphatic will not return to the capillary after the compression cycle is over because the many valves in tha collecting lymphatic block any backward flow.

Thus, any factor that causes compression of the lymphatic capillaries probably causes them to pump fluid in the same way that compression of the larger lymphatics causes pumping.

Flow of Lymph into the Lymphatic Capillary Despite Negative Pressure in the Interstitial Spaces. It has been difficult for many students of physiology to understand how fluid can flow out of the interstitial spaces into the lymphatic capillary in the face of the negative pressure in the interstitial spaces averaging −5 mm Hg, as was discussed in the previous chapter. However, a likely way that this occurs is the following: Every time a tissue is compressed, interstitial fluid pressure in the local area of compression rises temporarily to a positive value. This causes minute amounts of fluid to move into the lymphatics and subsequently to be pumped away from the tissues. Then, upon removal of the compression, the recoiling of the elastic structures in the tissues, particularly of the reticular structure of the tissue gel, creates suction in the tissue spaces. Thus, except during the momentary periods of compression, a negative pressure could in this way be maintained in the tissue spaces.

Summary of Factors That Determine Lymph Flow. From the above discussion one can see that

the two primary factors that determine lymph flow are the interstitial fluid pressure and the activity of the lymphatic pump. Therefore, one can state that, roughly, *the rate of lymph flow is determined by the product of interstitial fluid pressure and the activity of the lymphatic pump.*

Maximum Limit to the Rate of Lymph Flow. Figure 31–4 illustrates the relationship between interstitial free fluid pressure (P_T) and the rate of lymph flow. Note that at low interstitial fluid pressures of −6 to −7 mm Hg the lymph flow is very low. However, as the interstitial fluid pressure rises, the flow increases more than 20 times. Yet, once the pressure rises slightly greater than 0 mm Hg (atmospheric pressure), the flow reaches a plateau beyond which it will not rise any more even though the interstitial fluid pressure continues to rise.

There seem to be two major reasons why lymph flow reaches a maximum limit: (1) Once the tissues become edematous, the lymphatic capillaries also become tremendously dilated. This causes the flap valves between the endothelial cells of the capillaries to become separated from each other so that they are no longer reliable; therefore, the lymphatic capillary pump begins to fail. (2) The interstitial fluid pressure presses on the outside of the larger lymph channels to cause them to collapse; therefore, the input pressures at the tips of the lymphatic capillaries are opposed by a compression of the lymphatic walls of equal magnitude.

This maximum limit of lymph flow is of major significance because it demonstrates that most of the compensation to prevent edema caused by increased lymph flow occurs before the edema actually appears. That is, this mechanism prevents the development of edema *before* it occurs rather than responding after edema is present. It is only those persons whose abnormalities have gone beyond the limits of this lymphatic compensating mechanism who actually develop edema.

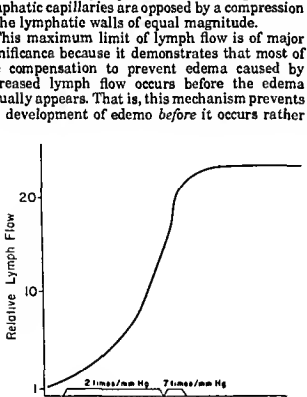


Figure 31–4. Relationship between interstitial fluid pressure and lymph flow. Note that lymph flow reaches a maximum as the interstitial pressure rises slightly above atmospheric pressure (0 mm Hg). (Courtesy of Drs. Harry Gibson and Aubrey Taylor.)

CONTROL OF INTERSTITIAL FLUID PROTEIN CONCENTRATION

REGULATION OF INTERSTITIAL FLUID PROTEIN BY LYMPHATIC PUMPING

Since protein continually leaks from the capillaries into the interstitial fluid spaces, it must also be removed continually, or otherwise the tissue colloid osmotic pressure will become so high that normal capillary dynamics can no longer continue. Therefore, by far the most important of all the lymphatic functions is the maintenance of low protein concentration in the interstitial fluid. The mechanism of this is the following:

As fluid leaks from the arterial ends of the capillaries into the interstitial spaces, only small quantities of protein accompany it, but then, as fluid is reabsorbed at the venous ends of the capillaries, most of the protein is left behind. Thus, excess quantities of *protein progressively accumulate in the interstitial fluid* and this in turn *increases the tissue colloid osmotic pressure.* The osmotic pressure then decreases reabsorption of fluid by the blood capillaries, thereby *promoting increased tissue fluid volume* and *less negativity of the interstitial fluid pressure.* The less negative pressure then allows the lymphatic pump to pump greater amounts of fluid into the lymphatic capillaries, and this fluid carries with it the excess protein that has accumulated. It is this continual *washout* of the protein that keeps the protein concentration at a low level in the interstitial fluid.

To summarize, an increase in tissue fluid protein increases the tissue fluid volume and therefore also the rate of lymph flow, and this washes the protein out of the tissue spaces, automatically returning the protein concentration to its normal low level.

The importance of this function of the lymphatics cannot be stressed too strongly, for *there is no other route besides the lymphatics through which excess proteins can return to the circulatory system.* If it were not for this continual removal of proteins from the tissues, the dynamics of fluid exchange at the blood capillaries would become so abnormal within only a few hours that life could no longer continue. There is certainly no other function of the lymphatics that can even approach this in importance.

CONTROL OF INTERSTITIAL FLUID PRESSURE

Interstitial fluid pressure is controlled at the same time that the interstitial fluid protein con-

centration is controlled. There are two aspects of this that we need specifically to discuss:

First, part of the control of interstitial fluid pressure results from increased lymph flow away from the tissue when the interstitial fluid pressure rises too high. This increased flow obviously decreases the interstitial fluid volume back toward normal and therefore also reduces the pressure back toward normal as well. However, this mechanism for removal of excess fluid from the interstitium is actually less important than the following mechanism:

Second, an increase in the interstitial fluid pressure causes an immediate imbalance of forces at the capillary membrane, because increased interstitial fluid pressure increases the tendency for water and small solute molecules to move through the capillary membrane from the interstitium into the capillaries. Therefore, excess fluid is absorbed into the capillaries until the forces across the capillary membrane again come to equilibrium. About −5 mm of the normal −5.3 mm Hg interstitial fluid pressure is caused by this balance of forces at the capillary membrane whereas only about −0.3 mm. Hg is caused by the continual drainage of fluid into the lymphatic vessels.

But, why is so much fluid absorbed directly into the capillaries? The answer to this is that the lymphatics continually remove protein from the interstitium. In this way they maintain a relatively low tissue protein concentration. Therefore, even though it is not the lymphatic removal of *fluid* from the tissues that maintains most of the negative interstitial fluid pressure, *the lymphatic removal of proteins is absolutely essential in causing the negative interstitial fluid pressure.* Thus, once again we see the tremendous importance of the lymphatic system for removing protein from the interstitial spaces, but less importance of the lymphatics for removing fluid. That is, if the proteins are continually removed so that the interstitial colloid osmotic pressure is kept at a low level, any excess fluid in the tissue spaces then can easily be removed by osmotic absorption directly into the capillaries.

Significance of Negative Interstitial Fluid Pressure as a Means for Holding the Body Tissues Together

In the past it has been assumed that the different tissues of the body are held together entirely by connective tissue fibers. However, at many places in the body, connective tissue fibers are absent. This occurs particularly at points where tissues slide over each other. Yet, even at these places, the tissues are held together by the negative in- [illegible] fluid pressure, which is actually a partial [illegible]m. When the tissues lose their negative pres- [illegible] fluid accumulates in the spaces, and the [illegible] known as *edema* occurs, which is discussed later in the chapter.

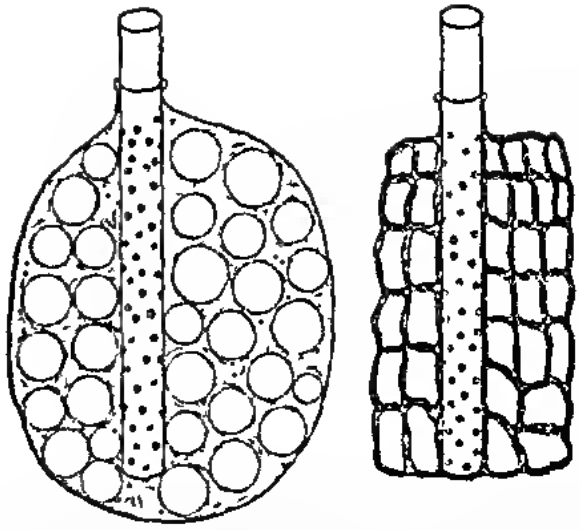


Figure 31–5. Physical model of the tissues constructed of a rubber bag, a perforated tube to simulate a capillary, balloons filled with water to simulate cells, and cotton between the balloons to simulate intercellular elements. *Left:* edematous state. *Right:* nonedematous state.

Significance of the Normally "Dry" State of the Interstitial Spaces. The normal tendency for the capillaries to absorb fluid from the interstitial spaces and thereby to create a partial vacuum causes all the minute structures of the interstitial spaces to be *compacted.* Figure 31–5 illustrates a physical model of the tissues constructed to illustrate this effect. To the left, positive pressure is present and excessive quantities of fluid are present in the "interstitial spaces." To the right, negative pressure has been applied through the perforated tube, which simulates a capillary, and the tissue elements are pulled tightly together. This represents a "dry" state; that is, no *excess* fluid is present besides that required simply to fill the crevices between the tissue elements.

EDEMA

Edema means the presence of excess interstitial fluid in the tissues. Referring again to Figure 31–5, one sees that the left-hand portion of the figure represents the edematous state, while the right-hand portion represents the nonedematous state.

Obviously, any factor that increases the interstitial fluid pressure high enough can cause excess interstitial fluid volume and thereby cause edema.

PRESSURE-VOLUME CURVE OF THE INTERSTITIAL FLUID SPACES

Figure 31–6 illustrates the average relationship between pressure and volume in the interstitial fluid spaces in the human body as extrapolated from measurements in the dog. The shape of the curve was determined in the following manner: A dog's leg was removed from its body and then perfused with concentrated dextran solution having a colloid osmotic pressure about twice that of normal plasma. This high colloid osmotic force inside the capillaries caused absorption of fluid

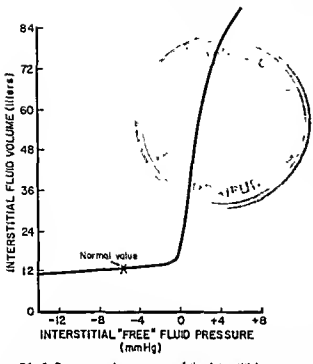


Figure 31–6. Pressure-volume curve of the interstitial spaces. (Extrapolated to the human being from data obtained in dogs.)

from the interstitial spaces and caused the weight of the leg to decrease. Measurement of this change in weight provided a means for measuring the decrease in interstitial fluid volume. Simultaneously, the free fluid pressure in the interstitial fluid spaces was measured using the implanted capsule method described in the preceding chapter. Later in the experiment, fluid having no colloid osmotic pressure was perfused through the limb, and this caused tremendous quantities of fluid to leak out of the capillaries into the interstitial spaces, thereby increasing the interstitial fluid volume. The curve of Figure 31–6 represents the average results obtained in experiments of this type.

The Slight Interstitial Fluid Volume Change in the Negative Pressure Range. One of the most significant features of the curve in Figure 31–6 is that so long as the interstitial fluid pressure remains in the negative range there is little change in interstitial fluid volume despite marked change in pressure. Therefore, detectable edema will not occur so long as the interstitial free fluid pressure remains negative. Indeed, in several hundred measurements of interstitial free fluid pressure made in experimental animals, no edema has ever been recorded in the presence of negative interstitial pressure.

Tremendous Increase in Interstitial Fluid Volume When the Interstitial Free Fluid Pressure Becomes Positive. Note in Figure 31–6 that just as soon as the interstitial free fluid pressure rises to equal atmospheric pressure (zero pressure), the slope of the pressure-volume curve suddenly changes and the volume increases precipitously. An additional increase in interstitial free fluid pressure of only 1 to 2 mm Hg now causes the interstitial fluid volume to increase several hundred per cent. Finally, at the very top of the figure, the skin begins to be stretched so that the volume now increases much less rapidly.

Tissue Space Compliance in Different Pressure Ranges. Another way to express the pressure-volume characteristics of the interstitial fluid spaces is in terms of *compliance*, which is defined as the *change in volume* for a given *change in pressure*. In the negative pressure range, the compliance of the interstitial spaces is slight, about 400 ml/mm Hg for the entire human body (as extrapolated from measurements in dogs). But, just as soon as the interstitial free fluid pressure rises into the positive range, this compliance increases tremendously, rising to about 10,000 ml/mm Hg. Thus, the compliance increases approximately 25-fold between the negative pressure range and the positive pressure range.

POSITIVE INTERSTITIAL FLUID PRESSURE AS THE PHYSICAL BASIS FOR EDEMA

After studying the pressure-volume curve of Figure 31–6, one can readily see that whenever the interstitial free fluid pressure rises above the surrounding atmospheric pressure, the tissue spaces begin to swell. Therefore, *the physical cause of edema is positive pressure (that is, supra-atmospheric pressure) in the interstitial fluid spaces.*

Degree of Edema in Relation to the Degree of Positive Pressure. The solid curve of Figure 31–7 is the same pressure-volume curve as that shown in Figure 31–6, but an edema scale has been added to the figure. One-plus edema means edema that is barely detectable, and 4+ edema means edema in which the limbs are swollen to diameters 1.5 to 2 times normal.

Note in Figure 31–7 that edema usually is not detectable in tissues until the interstitial fluid volume has risen to about 30 per cent above normal. And note also that the interstitial fluid volume increases to several hundred per cent above normal in seriously edematous tissues.

Stretch of the Tissue Spaces in Chronic Edema. If edema persists even for a few hours, and especially if

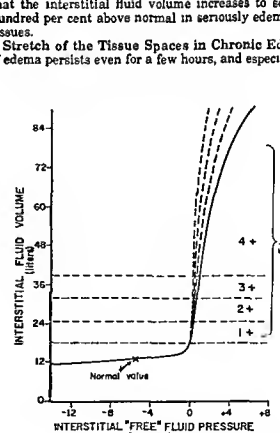


Figure 31–7. Relationship of edema to the pressure-volume curve of the interstitial spaces. The dashed curves show the eff of prolonged edema on the pressure-volume curve.

it persists for weeks, months, or years, the tissue spaces gradually become stretched As a result, the pressure-volume curve changes from the solid curve in Figure 31-7 to the progressively higher dashed curves. In other words, in chronic edema the tissue "bag" expands, which increases the ease with which the tissues can develop severe edema. Even a 1 to 2 mm Hg pressure rise above atmospheric pressure can cause 4+ edema once the tissue spaces have been stretched for many days This phenomenon of stretch of the tissues is called *delayed compliance* or *stress-relaxation of the tissue spaces*

The Phenomenon of "Pitting" Edema. If one presses a finger on the skin over an edematous area and then suddenly removes the finger, a small depression called a "pit" remains. Gradually, within 5 to 30 seconds, the pit disappears The cause of pitting is that edema fluid has been translocated away from the area beneath the pressure point The fluid simply flows through the tissue spaces to other tissue areas Then, when the finger is removed, 5 to 30 seconds is required for the fluid to flow back into the area from which it had been displaced

Nonpitting Edema. Occasionally, the fluid in seriously edematous tissues cannot be mobilized to other areas by pressure. The usual cause of this is coagulation of the protein, especially the fibrinogen, in the tissues For instance, in an infected or traumatized area, large quantities of fluid can collect, but coagulation of the protein entraps the fluid in the form of a gel and therefore prevents the fluid from being expulsed by pressure. Also, swelling of tissue cells, which occurs when cells are traumatized, diseased, or fail to receive adequate nutrition, can also cause nonpitting edema This type of edema is frequently also called *brawny* edema.

THE CONCEPT OF AN "EDEMA SAFETY FACTOR"

Safety Factor Caused by the Normal Negative Interstitial Free Fluid Pressure. One can readily see from Figures 31-6 and 31-7 that the interstitial fluid pressure must rise from the normal value of −5 3 to above 0 mm Hg before edema begins to appear Thus, there is a safety factor of 5 3 mm Hg caused by the normal negative interstitial free fluid pressure before edema appears.

Safety Factor Caused by Flow of Lymph from the Tissues. Another safety factor that helps to prevent edema is increased lymph flow When the interstitial free fluid pressure rises above the normal value of −5 3 mm Hg, lymph flow increases greatly, which removes a greater portion of the extra fluid entering the interstitial spaces And this obviously helps to prevent the development of edema

One can estimate that maximally increased lymph flow gives approximately a 7 mm Hg safety factor, for the extra lymph flow can carry away from the tissues approximately the extra amount of fluid that is formed by a 7 mm Hg excess in capillary pressure

Safety Factor Caused by Washout of Protein from the Interstitial Spaces. In addition to removal of fluid volume from the interstitial fluid spaces, increased lymph flow also washes out most of the proteins from the interstitial fluid spaces, decreasing the colloid osmotic pressure of the interstitial fluid from the normal value of 6 mm Hg down to about 1 mm Hg This provides another 5 mm Hg safety factor.

Total Safety Factor and Its Significance. Now, let us add all the above safety factors:

	mm Hg
Negative interstitial fluid pressure	5 3
Lymphatic flow	7 0
Lymphatic washout of proteins	5 0
TOTAL	17.3

Thus we find that a total safety factor of about 17 mm Hg is present to prevent edema. This means that the capillary pressure must rise about 17 mm Hg above its normal value of 17 mm Hg—that is, above 34 mm Hg—before edema begins to appear Or the plasma colloid osmotic pressure must fall from the normal level of 28 mm Hg to below 11 mm Hg before edema begins to appear. This explains why the normal human being does not become edematous until severe abnormalities occur in the circulatory system.

EDEMA RESULTING FROM ABNORMAL CAPILLARY DYNAMICS

From the discussion of capillary and interstitial fluid dynamics in the preceding and present chapters, it is already evident that several different abnormalities in these dynamics can increase the tissue pressure and in turn cause extracellular fluid edema. The different causes of extracellular fluid edema are:

Increased Capillary Pressure as a Cause of Edema. Figure 31-8A shows the effect of increased mean capillary pressure on the dynamics of fluid exchange at the capillary membrane. In this example, the mean capillary pressure is 41 mm Hg instead of the usual normal 17 mm Hg Consequently, enough fluid flows into the tissue spaces to raise the interstitial fluid pressure to +8 mm Hg This is far above the atmospheric pressure of 0 mm Hg, and therefore causes progressive enlargement of the tissue spaces, with tremendous expansion of the extracellular fluid volume

Causes of Increased Capillary Pressure. Increased capillary pressure can result from any clinical condition that causes either *venous obstruction* or *arteriolar dilatation*. Large venous blood clots frequently cause local areas of venous obstruction, which block the return of blood to the heart and promote edema in the tissues normally drained by the obstructed veins.

More frequently, however, capillary pressure is increased by obstruction of venous return due to *cardiac failure*, for, when the heart no longer pumps blood out of the veins with ease, blood dams up in the venous system The capillary pressure rises, and serious "cardiac edema" occurs. The dynamics of this type of edema are complicated, however, and were discussed in detail in Chapter 26

Decreased Plasma Proteins as a Cause of Edema. Figure 31-8B illustrates the abnormal dynamics that occur at the capillary membrane when the plasma protein concentration falls to an abnormally low value The major effect is a markedly lowered colloid osmotic pressure of the plasma. Consequently, the capillary pressure far overbalances the colloid osmotic pressure, causing fluid to collect in the tissue spaces, and the interstitial free fluid pressure rises As long as the pressure remains excessively elevated, the tissue spaces continually enlarge, with the edema becoming progressively worse

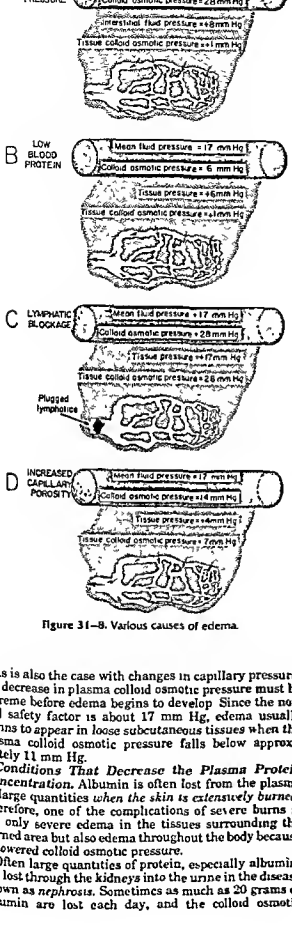


Figure 31–8. Various causes of edema.

As is also the case with changes in capillary pressure, the decrease in plasma colloid osmotic pressure must be extreme before edema begins to develop. Since the normal safety factor is about 17 mm Hg, edema usually begins to appear in loose subcutaneous tissues when the plasma colloid osmotic pressure falls below approximately 11 mm Hg.

Conditions That Decrease the Plasma Protein Concentration. Albumin is often lost from the plasma in large quantities *when the skin is extensively burned*. Therefore, one of the complications of severe burns is not only severe edema in the tissues surrounding the burned area but also edema throughout the body because of lowered colloid osmotic pressure.

Often large quantities of protein, especially albumin, are lost through the kidneys into the urine in the disease known as *nephrosis*. Sometimes as much as 20 grams of albumin are lost each day, and the colloid osmotic pressure of the plasma may fall to one-third normal or even less. This results in severe edema, and the edema itself is likely to contribute to death by means that are discussed later in the chapter.

Finally, persons who do not have sufficient protein in their diets are unable to form adequate quantities of plasma protein and, therefore, are likely to develop protein deficiency edema, which is called *nutritional edema*. This occurs frequently in famine areas, especially in some tribal areas of Africa.

Lymphatic Obstruction as a Cause of Edema. Even under normal conditions a small amount of protein leaks continually from the capillaries into the tissue spaces, and the only route by which the protein can be returned to the circulatory system is through the lymphatics. If the lymphatic drainage becomes blocked, more and more protein collects in the local tissue spaces until the concentration of the protein approaches the concentration of protein in the plasma. As shown in Figure 31–8C, the colloid osmotic pressure of the tissue fluids may theoretically rise to as high as 28 mm Hg, and, to balance this, fluid collects in the tissues until the interstitial fluid pressure rises to a value equal to the capillary pressure, about 17 mm Hg. Such elevated tissue pressure rapidly expands the tissue spaces, with resultant edema of the severest kind.

Causes of Lymphatic Obstruction. One of the most common causes of lymphatic obstruction is *filariasis*—that is, infection by nematodes of the superfamily Filarioidea. The disease is widespread in the tropics, where larvae (microfilariae) are transmitted to human hosts by mosquitoes. The larvae pass out of the capillaries into the interstitial fluid and then by way of the lymph into the lymph nodes. Subsequent inflammatory reactions progressively obstruct the lymphatic channels of these nodes with scar tissue. After several years, the lymphatic drainage from one of the peripheral parts of the body may become almost totally occluded. Thus, a leg can swell to such a size that it might weigh as much as all the remainder of the body. Because of this extreme degree of edema, the swollen condition is frequently called *elephantiasis*. An interesting type of elephantiasis occasionally occurs in the scrotum, which has been known to enlarge so much that the person must carry it in a wheelbarrow in order to move about.

Lymphatic obstruction also occurs following operations for removal of cancerous tissue. Because the lymph nodes draining a cancerous area of the body must be removed in order to prevent possible spread of the cancer, the return of lymph to the circulatory system from that area will be blocked. Occasionally a radical mastectomy for removal of a cancerous breast causes the corresponding arm to swell to as much as twice its normal size, but usually the swelling regresses during the following two to three months as new lymph channels develop.

Increased Permeability of the Capillaries as a Cause of Edema. Figure 31–8D illustrates a capillary whose membrane has become so permeable that even protein molecules pass from the plasma into the interstitial spaces with ease. The protein content of the plasma decreases while that of the interstitial spaces increases. In the example of the figure the tissue pressure rises to +7 mm Hg to balance the changes in plasma and tissue colloid osmotic pressure occasioned by the leakage of protein. The elevated interstitial free fluid pressure in turn causes progressive edema.

Cause of Increased Capillary Permeability. Cap-

illaries become excessively permeable when any factor destroys the integrity of the capillary endothelium. Burns are a frequent cause of increased permeability of the capillaries because overheated capillaries become friable, and their pores enlarge. Allergic reactions also frequently cause the release of histamine or various polypeptides that damage the capillary membranes and cause increased permeability.

EDEMA CAUSED BY KIDNEY RETENTION OF FLUID

When the kidney fails to excrete adequate quantities of urine, and the person continues to drink normal amounts of water and ingest normal amounts of electrolytes, the total amount of extracellular fluid in the body increases progressively. This fluid is absorbed from the gut into the blood and elevates the capillary pressure This in turn causes most of the fluid to pass into the interstitial fluid spaces, elevating the interstitial fluid pressure as well. Therefore, simple retention of fluid by the kidneys can result in extensive edema

THE PRESENCE AND IMPORTANCE OF GEL IN THE INTERSTITIAL SPACES

Up to this point we have discussed interstitial fluid as if it were entirely in a mobile, "free" state. However, in the normal interstitial spaces, the interstitial fluid is entrapped in a *gel-matrix* comprised of large *proteoglycan molecules* (also called *mucopolysaccharides)*. These molecules generally have molecular weights greater than one million, are folded into odd shapes, and are entangled among each other, which is what causes the gel-like nature of the normal interstitial fluid. The widths of the spaces between the molecules are usually only 20 to 40 nanometers, which is so small that water molecules and dissolved substances in the interstitial fluids *cannot flow with ease* through this gel-matrix. Therefore, the interstitial fluid in normal tissues is relatively immobilized.

Even though the fluid in the interstitial gel cannot "flow" easily from one part of the interstitium to another, the individual molecules still do move by random kinetic motion as rapidly as ever Furthermore, since most of these molecules have diameters 20 or more times less than the sizes of the spaces between the proteoglycan molecules, they can still move by the process of *diffusion* through the interstitium more than 95 per cent as effectively as in free fluid. Therefore, nutrients can diffuse from the capillaries to the cells almost equally as well through the gel as through free fluid.

Importance of the Gel-Matrix. There are several important advantages of having a gel-matrix in the interstitium. Some of these are:

(1) The proteoglycan molecules act as a "filler" to hold the cells apart. This creates large enough spaces for fluid and nutrients to diffuse from the capillaries to those cells that are located some distance away from the capillaries.

(2) Since the fluid in the tissue spaces is mainly immobilized by the gel, this prevents fluid from flowing through the tissue spaces from the upper part of the body into the lower body. Otherwise all of the interstitial fluid (16 per cent of the total body weight) would flow within minutes into the tissue spaces of the legs and feet.

(3) The proteoglycan meshwork not only immobilizes fluid but also immobilizes bacteria and keeps them from spreading through the tissues.

Relation of Edema Fluid to the Gel. When excessive amounts of fluid begin to accumulate in the interstitial spaces, the gel at first entraps this additional fluid as well, and the entire gel-matrix of the interstitium swells. However, once the gel has swollen more than 30 to 50 per cent, the entanglements of the proteoglycan molecules begin to break up, and then *free fluid spaces* develop all through the interstitium. As still more fluid accumulates, the free fluid spaces become so large that they coalesce to form large free fluid channels in the tissues. Once this has occurred, fluid then flows freely through the tissues, 10,000 to 250,000 times as easily as in normal tissues.

Figure 31–9 illustrates the volume relationship

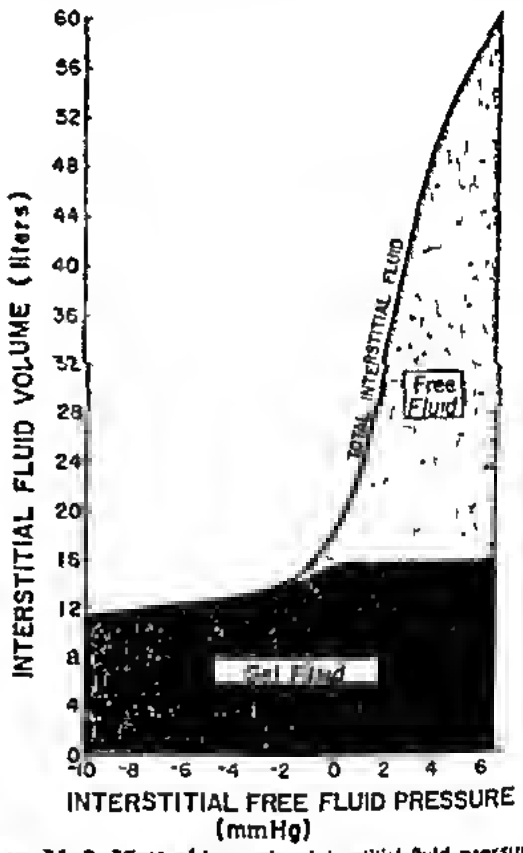


Figure 31–9. Effect of increasing interstitial fluid pressure on the volumes of total interstitial fluid, gel fluid, and free fluid. Note that significant amounts of free fluid occur only when the interstitial free fluid pressure becomes positive. (Modified from Guyton, Granger, and Taylor: *Physiol Rev.*, 51 527, 1971.)

between free interstitial fluid, gel fluid, and total interstitial fluid in both the nonedematous and the edematous states. Note that under normal conditions, when the interstitial free fluid pressure is in its normal negative pressure range, there is almost imperceptible free fluid in the tissues. Instead, almost all of the fluid is in the gel phase and it is also highly immobile. On the other hand, as the interstitial free fluid pressure rises and as the state of edema approaches, the gel swells some 30 to 50 per cent, after which it swells no more. With still more increase in interstitial free fluid pressure, all the additional edema fluid that accumulates is free fluid that is highly mobile through the tissue spaces. It is this high degree of mobility that causes the edema to be of the pitting type, as was explained earlier in the chapter.

Relationship of Interstitial Fluid Gel to the Regulation of Interstitial Fluid Volume. Since approximately 16 per cent of the average tissue is composed of interstitial fluid and normally almost all of it is in the gel state, one can derive the following concept for the regulation of interstitial fluid volume. The mechanism previously described for creating a negative pressure in the tissue spaces is actually a "drying" mechanism that always attempts to remove any *free* fluid that appears in the tissues. Thus, normally, essentially all free fluid is removed as rapidly as it is formed, leaving in the normal tissues only the gel, which still constitutes about 16 per cent of the tissue volume.

The question remains, though, why does this drying mechanism remove only a small amount of fluid from the gel? The answer to this has two components: First, the fine reticular filaments of the gel are composed of hyaluronic acid molecules that are coiled like springs and are compressed against each other. Therefore, the elastic forces of these molecules prevent excessive compression in the same way that cotton fibers in cloth fabrics prevent compression beyond a certain point. Second, the gel has a slight amount of osmotic pressure caused by the Donnan equilibrium effect; that is, the gel reticulum has negative electrostatic charges that hold small, mobile, positive ions—mainly sodium ions—within the gel. These ions in turn cause osmosis of water into the gel.

The quantity of mucopolysaccharides in the tissue gel is sufficient to give an osmotic absorptive pressure in the gel that is calculated to be about 2 mm Hg. The elastic recoil of the hyaluronic acid "springs" gives approximately another 4 mm Hg of recoil force, making a total of about 6 mm Hg to resist dehydration caused by the −5.3 mm Hg pressure in the free fluid of the tissue spaces.

PULMONARY INTERSTITIAL FLUID DYNAMICS

The dynamics of the pulmonary interstitial fluid are essentially the same as those for fluid in the peripheral tissues except for the following important quantitative differences:

1. The pulmonary capillary pressure is very low in comparison with the systemic capillary pressure, approximately 7 mm Hg in comparison with 17 mm Hg.

2. The interstitial free fluid pressure in the lung interstitium has been measured to be −8 mm Hg in comparison with −5 mm Hg in subcutaneous tissue.

3. The pulmonary capillaries are relatively leaky to protein molecules so that the protein concentration of lymph leaving the lungs is relatively high, averaging about 4 grams per cent instead of 2 grams per cent in the peripheral tissues.

4. The rate of lymph flow from the lungs is also very high, mainly because of the continuous pumping motion of the lungs.

5. The interstitial spaces of the alveolar portions of the lungs are very narrow, represented by the minute spaces between the capillary endothelium and the alveolar epithelium.

6. The alveolar epithelia are not strong enough to resist very much positive pressure. They are probably ruptured by any positive pressure in the interstitial spaces greater than atmospheric pressure (0 mm Hg), which allows dumping of fluid from the interstitial spaces into the alveoli.

Now let us see how these quantitative differences affect pulmonary fluid dynamics.

Interrelationship Between Interstitial Fluid Pressure and Other Pressures in the Lung. Figure 31–10 illustrates a pulmonary capillary, a pulmonary alveolus, and a lymphatic capillary draining the interstitial space between the capillary and the alveolus. Note the balance of forces at the capillary membrane as follows:

	mm Hg
Forces tending to cause movement of fluid outward from the capillaries and into the pulmonary interstitium:	
Capillary pressure	7
Interstitial fluid colloid osmotic pressure	14
Negative interstitial fluid pressure	8
TOTAL OUTWARD FORCE	29
Forces tending to cause absorption of fluid into the capillaries:	
Plasma colloid osmotic pressure	28
TOTAL INWARD FORCE	28

Note that the normal outward forces are slightly greater than the inward forces. The *net mean filtration pressure* at the pulmonary capillary membrane can be calculated as follows:

	mm Hg
Total outward force	+29
Total inward force	−28
NET MEAN FILTRATION PRESSURE	+1

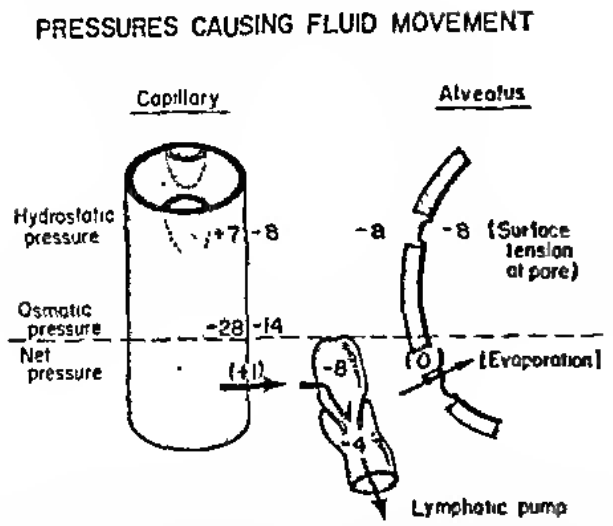


Figure 31–10. Hydrostatic and osmotic forces at the capillary (left) and alveolar membrane (right) of the lungs. Also shown is a lymphatic (center) that pumps fluid from the pulmonary interstitial spaces. (Modified from Guyton, Taylor, and Granger: Dynamics and Control of the Body Fluids. Philadelphia, W. B. Saunders Company, 1975.)

This net filtration pressure causes a slight continual flow of fluid from the pulmonary capillaries into the interstitial spaces, and, except for a small amount that evaporates in the alveoli, this fluid is pumped back to the circulation through the pulmonary lymphatic system.

Fluid Exchange at the Pulmonary Alveolar Membrane; The Mechanism for Keeping the Alveoli "Dry." One of the most important problems in lung function is to understand why the alveoli do not fill up with fluid. One's first impulse is to say that the alveolar epithelium keeps fluid from leaking out of the interstitial spaces into the alveoli. However, this is not true, because there are always a small number of fairly large openings between the alveolar epithelial cells through which even large protein molecules as well as large quantities of water and electrolytes can pass.

However, if one remembers that the pulmonary capillaries and the pulmonary lymphatic system normally maintain a negative pressure in the interstitial spaces, then it is clear that whenever extra fluid appears in the alveoli this will simply be sucked mechanically into the lung interstitium through the small holes between the alveolar epithelial cells. Then the excess fluid is either sucked into the pulmonary lymphatics or is absorbed into the pulmonary capillaries. Thus, under normal conditions the alveoli are kept in a "dry" state except for a small amount of fluid that seeps from the epithelium onto the lining surfaces of the alveoli to keep them moist.

EDEMA

Pulmonary edema occurs in the same way that it occurs elsewhere in the body. Any factor that causes the pulmonary interstitial fluid pressure to rise from t negative range into the positive range will cause sudd filling of the pulmonary interstitial spaces and, in m severe cases, even the alveoli with large amounts of f fluid.

The usual causes of pulmonary edema are:

(1) Left heart failure or mitral valvular disease w consequent great increase in pulmonary capillary pr sure and flooding of the interstitial spaces.

(2) Damage to the pulmonary capillary membr caused by breathing noxious substances such as chlor gas or sulfur dioxide gas.

(3) Decrease in plasma colloid osmotic pressure t low enough level that fluids transude from the bl into the pulmonary interstitial spaces (occurs o rarely).

Pulmonary "Interstitial Fluid" Edema Vers Pulmonary "Alveolar" Edema. The interstitial f volume of the lungs usually cannot increase more th about 50 per cent (representing less than 100 milliit of fluid) before the alveolar epithelial membranes r ture and fluid begins to pour from the interstitial spa into the alveoli. The cause of this is simply the v slight tensional strength of the pulmonary alveo epithelium; that is, any positive pressure in the int stitial fluid spaces seems to cause immediate rupture this epithelium.

Therefore, except in the mildest cases of pulmon edema, edema fluid always enters the alveoli; if t edema becomes severe enough, it can cause death suffocation, as has already been discussed in Chapt 26 and 27.

The "Pulmonary Edema Safety Factor." All factors that tend to prevent edema in the periphe tissues also tend to prevent edema in the lungs. T is, before positive interstitial fluid pressure can oc and cause edema all the following factors must overcome: (1) the normal negativity of the interstit fluid pressure of the lungs, (2) the lymphatic pump of fluid out of the interstitial spaces, and (3) the creased osmosis of fluid into the pulmonary capilla caused by decreased protein in the interstitial fluid w the lymph flow increases.

In experiments in animals it has been found that pulmonary capillary pressure normally must rise t value at least equal to the plasma colloid osmotic pr sure before significant pulmonary edema will occ Thus, in the human being, who normally has a plas colloid osmotic pressure of 28 mm Hg, one can pred that the pulmonary capillary pressure must rise fr the normal level of 7 mm Hg to over 28 mm Hg to ca pulmonary edema, giving a safety factor against p monary edema of about 21 mm Hg.

Safety Factor in Chronic Conditions. When pulmonary capillary pressure remains elevated chro cally (for at least two weeks), the lungs become e more resistant to pulmonary edema because the lym vessels expand greatly, increasing their capability carrying fluid away from the interstitial spaces as m as tenfold. Therefore, a patient with chronic mit stenosis frequently has a pulmonary capillary pressu as great as 40 to 45 mm Hg without having signific pulmonary edema.

Thus, in chronic pulmonary edema, the safety fac against pulmonary edema can rise to as high as 35 40 mm Hg, in comparison with a normal value of mm Hg under acute conditions.

Rapidity of Death in Acute Pulmonary Edema. When the pulmonary capillary pressure does rise even slightly above the safety factor level, lethal pulmonary edema can occur within hours, and within 20 to 30 minutes if it is as much as 25 to 30 mm Hg above the safety factor level. Thus, in acute left heart failure, in which the pulmonary capillary pressure occasionally rises to as high as 50 mm Hg, death frequently ensues in less than one-half hour from acute pulmonary edema.

REFERENCES

Aukland, K , and Nicolaysen, G Interstitial fluid volume Local regulatory mechanisms *Physiol Rev , 51* 556, 1981

Bassingthwaighte, J B , and Goresky, C A. Modeling In the analysis of solute and water exchange in the microvasculature *In* Renkin, E M., and Michel, C C (eds) Handbook of Physiology Sec 2, Vol IV Bethesda, American Physiological Society, 1984, p 549

Bengis, R G , and Guyton, A C Some pressure and fluid dynamic characteristics of the canine epidural space *Am J Physiol , 232* H255, 1977

Bert, J L., and Pearce, R H The interstitium and microvascular exchange *In* Renkin, E M , and Michel, C C (eds) Handbook of Physiology Sec 2, Vol IV Bethesda, American Physiological Society, 1984, p. 521.

Brace, R A , *et al* Time course of lymph protein concentration in the dog Microvasc Res , *14* 243, 1977

Brigham, K L. Lung lymph composition and flow in experimental pulmonary edema. *In* Fishman, A P , and Renkin, E M (eds) Pulmonary Edema Baltimore, Waverly Press, 1979, p 161

Casley Smith, J. R. Lymph and lymphatics. *In* Kaley, G., and Altura, B M (eds) Microcirculation Vol 1 Baltimore, University Park Press, 1977, p 423

Clodius, L. (ed) Lymphedema Stuttgart, Thieme, 1977

Curry, F-R. E Mechanics and thermodynamics of transcapillary exchange *In* Renkin, E M , and Michel, C C (eds) Handbook of Physiology Sec 2, Vol IV Bethesda, American Physiological Society, 1984, p 309

Drinker, C K. The Lymphatic System Stanford, Stanford University Press, 1942.

Fishman, A P , and Pietra, G G Hemodynamic pulmonary edema *In* Fishman, A. P , and Renkin, E M (eds) Pulmonary Edema Baltimore, Waverly Press, 1979, p 79

Fishman, A P , and Renkin, E M (eds.) Pulmonary Edema. Baltimore, American Physiological Society, 1979

Gaar, K A , Jr , *et al* Effects of capillary pressure and plasma protein on development of pulmonary edema *Am J Physiol , 213* 79, 1967

Gaar, K. A., Jr , *et al* Effect of lung edema on pulmonary capillary pressure. *Am J Physiol , 216* 1370, 1969

Gross, J F., and Popel, A (eds) Mathematics of Microcirculation Phenomena New York, Raven Press, 1980

Guyton, A C ' Interstitial fluid pressure II Pressure-volume curves of interstitial space *Circ. Res , 16* 452, 1965

Guyton, A C Interstitial fluid pressure-volume relationships and their regulation. *In* Wolstenholme, G E W , and Knight, J (eds). CIBA Foundation Symposium on Circulatory and Respiratory Mass Transport London, J & A Churchill, 1969, p 4

Guyton, A C Introduction to Part I Pulmonary alveolar-capillary interface and interstitium *In* Fishman, A P , and Hecht, H H (eds) The Pulmonary Circulation and Interstitial Space Chicago, University of Chicago Press, 1969, p 3

Guyton, A C , and Coleman, T. G Regulation of interstitial fluid volume and pressure *Ann N Y Acad Sci , 150* 537, 1968.

Guyton, A. C., and Lindsey, A. W., Effect of elevated left atrial pressure and decreased plasma protein concentration on the development of pulmonary edema. *Circ Res , 7* 649, 1959

Guyton, A C , *et al* ' Interstitial fluid pressure *Physiol Rev , 51* 527, 1971

Guyton, A C , *et al* : Dynamics and Control of the Body Fluids. Philadelphia, W. B Saunders Co , 1975

Guyton, A C , *et al* : Circulatory Physiology II Dynamics and Control of the Body Fluids Philadelphia, W. B Saunders Co , 1975

Guyton, A C., *et al*.' Dynamics of subatmospheric pressure in the pulmonary interstitial fluid *In* Lung Liquids. CIBA Symposium Amsterdam, Elsevier/North-Holland, 1976, p 77

Guyton, A C , *et al* . Forces governing water movement in the lung *In* Fishman, A P., and Renkin, E M (eds) Pulmonary Edema Baltimore, Waverly Press, 1979, p 65

Kennedy, J F Proteoglycans. Biological and chemical aspects of human life New York, Elsevier Scientific Publishing Co , 1979

Kinmonth, J G ' The Lymphatics Surgery, Lymphography and Diseases of the Chyle and Lymph Systems Baltimore, E Arnold, 1982

Leak, L V , and Burke, J F Ultrastructural studies on the lymphatic anchoring of filaments *J Cell Biol , 36* 129, 1968

Lennarz, W J (ed) The Biochemistry of Glycoproteins and Proteoglycans New York, Plenum Press, 1979

Mayerson, H S The physiologic importance of lymph *In* Hamilton, W F (ed) Handbook of Physiology Sec 2 Vol 2 Baltimore, Williams & Wilkins, 1963, p 1035

Meyer, B J , *et al* Interstitial fluid pressure V Negative pressure in the lungs. *Circ Res , 22* 263, 1968

Nicoll, P A , and Taylor, A E Lymph formation and flow *Annu Rev Physiol , 39* 73, 1977

Nieuwenhuis, P (ed.). In Vivo Immunology Histophysiology of the Lymphoid System New York, Plenum Publishing Corp , 1982

Notter, R H , and Finkelstem, J N Pulmonary surfactant. An interdisciplinary approach *J Appl Physiol , 57* 1613, 1984

Parker, J C , *et al* Pulmonary interstitial and capillary pressures estimated from mitra alveolar fluid pressures *J Appl Physiol , 44*(2) 267, 1978

Parker, J C , *et al* Pulmonary transcapillary exchange and pulmonary edema *In* Guyton, A C , and Young, D B (eds] International Review of Physiology Cardiovascular Physiology III Vol 18 Baltimore, University Park Press, 1979, p 261

Renkin, E M Lymph as a measure of the composition of interstitial fluid *In* Fishman, A. P , and Renkin, E M (eds) Pulmonary Edema. Baltimore, Waverly Press, 1979, p 145

Robin, E D Permeability pulmonary edema *In* Fishman, A. P , and Renkin, E M (eds.) Pulmonary Edema Baltimore, Waverly Press, 1979, p 217

Schneeberger, E. E. Barrier function of intracellular junctions in adult and fetal lungs *In* Fishman, A P , and Renkin, E M (eds) Pulmonary Edema Baltimore, Waverly Press, 1979, p 21

Staub, N C (ed) Lung Water and Solute Exchange New York, Marcel Dekker, 1978

Staub, N C. Pathways for fluid and solute fluxes in pulmonary edema *In* Fishman, A P , and Renkin, E M (eds) Pulmonary Edema. Baltimore, Waverly Press, 1979, p 113

Staub, N C. Pulmonary edema due to increased microvascular permeability. *Annu. Rev Med , 32*.291, 1981

Tepletz, C Pulmonary cellular and interstitial edema *In* Fishman, A P , and Renkin, E. M (eds.). Pulmonary Edema Baltimore, Waverly Press, 1979, p 97.

Visscher, M B , *et al* ' The physiology and pharmacology of lung edema. *Pharm Rev , 8*.389, 1956

Weibel, E. R., and Bachofen, H , Structural design of the alveolar septum and fluid exchange. *In Fishman, A. P. and* Renkin, E M (eds) Pulmonary Edema. Baltimore, Waverly Press, 1979, p 1

Yoffey, J M , and Courtice, F. C. (eds.). Lymphatics, Lymph and Lymphomyeloid Complex. New York, Academic Press, 1970.

32

The Special Fluid Systems of the Body—Cerebrospinal, Ocular, Pleural, Pericardial, Peritoneal, and Synovial

Several special fluid systems exist in the body, each performing functions peculiar to itself. For instance, the cerebrospinal fluid supports the brain in the cranial vault, the intraocular fluid maintains distension of the eyeballs so that the optical dimensions of the eye remain constant, and the potential spaces, such as the pleural and pericardial spaces, provide lubricated chambers in which the internal organs can move. All these fluid systems have characteristics that are similar to each other and that are also similar to those of the interstitial fluid system. However, they are also sufficiently different that they require special consideration.

THE CEREBROSPINAL FLUID SYSTEM

The entire cavity enclosing the brain and spinal cord has a volume of approximately 1650 ml, and about 150 ml of this volume is occupied by cerebrospinal fluid. This fluid, as shown in Figure 32–1, is found in the ventricles of the brain, in the cisterns around the brain, and in the subarachnoid space around both the brain and the spinal cord. All these chambers are connected with each other, and the pressure of the fluid is regulated at a constant level.

CUSHIONING FUNCTION OF THE CEREBROSPINAL FLUID

A major function of the cerebrospinal fluid is to cushion the brain within its solid vault. Fortunately, the [illegible] and the cerebrospinal fluid have approximately the same specific gravity, so that the brain simply floats in the fluid. Therefore, a blow on the head moves the entire brain simultaneously, causing no one portion of the brain to be momentarily contorted by the blow.

Contrecoup. When a blow to the head is extremely severe, it usually does not damage the brain on the side of the head where the blow is struck, but, instead, on the opposite side. This phenomenon is known as "contrecoup," and the reason for this effect is the following. When the blow is struck, the fluid on the struck side is so incompressible that, as the skull moves, the fluid pushes the brain at the same time. However, on the opposite side, the sudden movement of the skull causes it to pull away from the brain momentarily, creating for

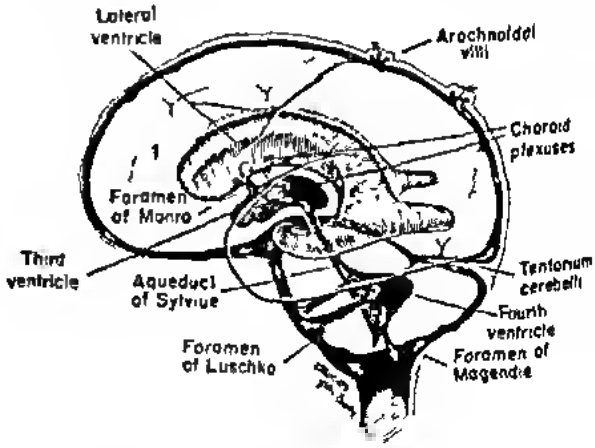


Figure 32–1. Pathway of cerebrospinal fluid flow from the choroid plexuses in the lateral ventricles to the arachnoidal villi protruding into the dural sinuses.

a short instant a vacuum in the cranial vault at this point. Then, when the skull is no longer being accelerated by the blow, the vacuum suddenly collapses and the brain strikes the inner surface of the skull. Because of this effect, the damage to the brain of a boxer usually does not occur in the frontal regions but, instead, in the occipital regions.

FORMATION, FLOW, AND ABSORPTION OF CEREBROSPINAL FLUID

Cerebrospinal fluid is formed at a rate of approximately 500 ml each day, which is about three times as much as the total volume of fluid in the entire cerebrospinal fluid system. Probably two thirds or more of this fluid originates as a secretion from the choroid plexus in each of the four ventricles, mainly in the two lateral ventricles. Additional amounts of fluid are secreted by all the ependymal surfaces of the ventricles, and a small amount comes from the brain itself through the perivascular spaces that surround the blood vessels entering the brain.

Figure 32–1 illustrates the main channel of fluid flow from the choroid plexuses and then through the cerebrospinal fluid system. The fluid secreted in the lateral ventricles passes through the *foramina of Monro* into the third ventricle, then combines with that secreted in the third ventricle, and finally passes along the *aqueduct of Sylvius* into the fourth ventricle where a small amount of additional fluid is added It then passes out of the fourth ventricle through three small openings, two lateral *foramina of Luschka* and a midline *foramen of Magendie*, entering the *cisterna magna*, a large fluid space that lies behind the medulla and beneath the cerebellum. The cisterna magna is continuous with the *subarachnoid space* that surrounds the entire brain and spinal cord, and almost all the cerebrospinal fluid flows upward through this space toward the cerebrum. But before it can reach the cerebrum it must first flow through the small *tentorial opening* around the mesencephalon where the flow is sometimes impeded From the cerebral subarachnoid spaces, the fluid flows into *arachnoidal villi* that project into the large sagittal venous sinus and to a lesser extent into other venous sinuses. Finally, the fluid empties into the venous blood through the surfaces of these villi.

Secretion by the Choroid Plexus. The choroid plexus, which is illustrated in Figure 32–2, is a cauliflower-like growth of blood vessels covered by a thin layer of epithelial cells. This plexus projects into (a) the temporal horn of each lateral ventricle, (b) the posterior portion of the third ventricle, and (c) the roof of the fourth ventricle.

The secretion of fluid by the choroid plexus depends mainly on active transport of sodium ions through the epithelial cells that line the outsides of the plexus. The sodium ions in turn pull along large amounts of chloride ions as well because the positive charge of the sodium ion attracts the chloride ion's negative charge. Finally, the two of these together increase the quantity of osmotically active substances in the cerebrospinal fluid, which then causes almost immediate osmosis of water through the membrane, thus providing the fluid of the secretion. Less important transport processes move small amounts of glucose into the cerebrospinal fluid and both potassium and bicarbonate ions out of the cerebrospinal fluid into the capillaries. Therefore, the

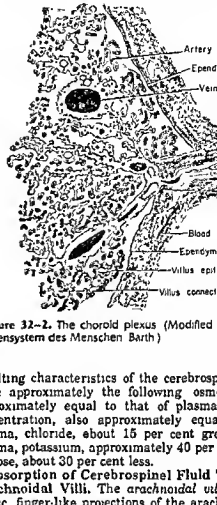


Figure 32–2. The choroid plexus (Modified from Clara Das Nervensystem des Menschen Barth)

resulting characteristics of the cerebrospinal fluid become approximately the following osmotic pressure, approximately equal to that of plasma, sodium ion concentration, also approximately equal to that of plasma, chloride, about 15 per cent greater than in plasma, potassium, approximately 40 per cent less; and glucose, about 30 per cent less.

Absorption of Cerebrospinal Fluid Through the Arachnoidal Villi. The *arachnoidal villi* are microscopic, finger-like projections of the arachnoidal membrana through the walls of the venous sinuses. Large conglomerates of these villi are usually found together and form macroscopic structures called *arachnoidal granulations* that can be seen protruding into the sinuses The endothelial cells covering the villi have been shown by electron microscopy to have large vesicular holes directly through the bodies of the cells. These are large enough to allow relatively free flow of cerebrospinal fluid, protein molecules, and even particles as large as red blood cells into the venous blood.

The Perivascular Spaces and Cerebrospinal Fluid. The blood vessels entering the substance of the brain pass first along the surface of the brain and then penetrate inward, carrying a layer of *pia mater* with them, as shown in Figure 32–3. The pia is only loosely adherent to the vessels, so that a space, the *perivascular space*, exists between it and each vessel. Perivascular spaces follow both the arteries and the veins into the brain as far as the arterioles and venules but not to the capillaries.

The Lymphatic Function of the Perivascular Spaces. As is true elsewhere in the body, a small amount of protein leaks out of the parenchymal capillaries into the interstitial spaces of the brain, and since no true lymphatics are present in brain tissue, this protein leaves the tissue mainly through the perivascular spaces but partly also by direct diffusion throug the pia mater into the subarachnoid spaces. On reachin the subarachnoid spaces, the protein flows along with the cerebrospinal fluid to be absorbed through the *arachnoidal villi* into the cerebral veins Therefore, the

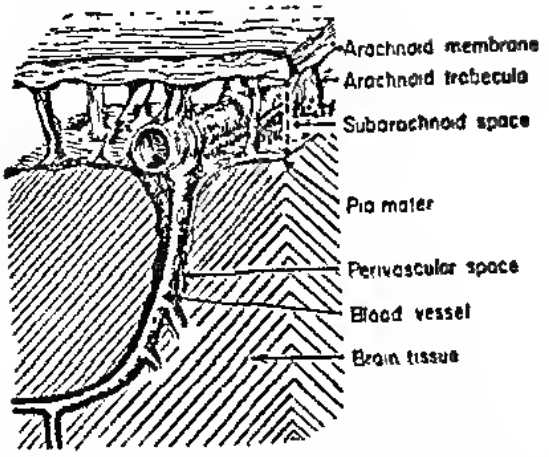


Figure 32–3. Drainage of the perivascular spaces into the subarachnoid space (From Ranson and Clark. Anatomy of the Nervous System. Philadelphia, W. B. Saunders Company, 1959.)

perivascular spaces, in effect, are a modified lymphatic system for the brain.

In addition to transporting fluid and proteins, the perivascular spaces also transport extraneous particulate matter from the brain into the subarachnoid space. For instance, whenever infection occurs in the brain, dead white blood cells are carried away through the perivascular spaces.

CEREBROSPINAL FLUID PRESSURE

The normal pressure in the cerebrospinal fluid system when one is lying in a horizontal position averages 130 mm water (10 mm Hg), though this may be as low as 70 mm water or as high as 180 mm water even in the normal person. These values are considerably more positive than the −5.3 mm Hg pressure in the interstitial spaces of the subcutaneous tissue.

Regulation of Cerebrospinal Fluid Pressure. The cerebrospinal fluid pressure is regulated almost entirely by absorption of the fluid through the arachnoidal villi. The reason for this is that the rate of cerebrospinal fluid formation is very constant, so that this is rarely a factor in pressure control. On the other hand, the villi act as valves that allow the fluid and its contents to flow readily into the venous blood of the sinuses while not allowing the blood to flow backward in the opposite direction. This valve action of the villi allows cerebrospinal fluid to flow into the blood when its pressure is 1.5 mm Hg greater than the pressure of the blood in the sinuses. Then, as the cerebrospinal fluid pressure rises still higher, the valves open very widely so that, under normal conditions, the pressure almost never rises more than a few millimeters Hg greater than the pressure in the venous sinuses.

On the other hand, in disease states the villi sometimes become blocked by large particulate matter or by fibrosis, which can cause very high cerebrospinal fluid pressure, as we shall discuss later.

Cerebrospinal Fluid Pressure in Pathological Conditions of the Brain. Often a large *brain tumor* elevates the cerebrospinal fluid pressure by decreasing the rate of absorption of fluid. For instance, if the tumor is above the tentorium and becomes so large that it compresses the brain downward, the upward flow of fluid through the tentorial opening may become blocked and the absorption of fluid greatly curtailed. As a resu the cerebrospinal fluid pressure can rise to as high 500 mm water (37 mm Hg) or more.

The pressure also rises considerably when *hemorrha* or *infection* occurs in the cranial vault. In both of the conditions, large numbers of cells suddenly appear the cerebrospinal fluid, and these can cause serio blockage of the small channels for absorption throu the arachnoidal villi. This sometimes elevates the cer brospinal fluid pressure to as high as 400 to 600 m water.

Occasionally babies are born with high cerebrospin fluid pressure. This is usually caused by abnormal high resistance to fluid reabsorption through the arac noidal villi, resulting either from too few arachnoid villi or villi with abnormal absorptive properties. Th is discussed later in connection with *hydrocephalus*.

Measurement of Cerebrospinal Fluid Pressur The usual procedure for measuring cerebrospinal flu pressure is the following: First, the subject lies exact horizontally on his side so that the spinal fluid pressu is equal to the pressure in the cranial vault. A spin needle is then inserted into the lumbar spinal can below the lower end of the cord and is connected to glass tube. The spinal fluid is allowed to rise in the tu as high as it will. If it rises to a level 136 mm above t level of the needle, the pressure is said to be 136 m water pressure or, dividing this by 13.6, which is t specific gravity of mercury, about 10 mm Hg pressure.

Effect of High Cerebrospinal Fluid Pressure the Optic Disc—Papilledema. Anatomically, the du of the brain extends as a sheath around the optic ner and then connects with the sclera of the eye. When t pressure rises in the cerebrospinal fluid system, it al rises in the optic nerve sheath. The retinal artery a vein pierce this sheath a few millimeters behind the e and then pass with the optic nerve into the eye itse The high pressure in the optic sheath impedes the fl of blood in the retinal vein, thereby also increasing t retinal capillary pressure throughout the eye, whi results in retinal edema. The tissues of the optic di are much more distensible than those of the remaind of the retina, so that the disc becomes far more edem tous than the remainder of the retina and swells in the cavity of the eye. The swelling of the disc, whi can be observed with an opthalmoscope, is called *pap ledema*, and neurologists can estimate the cerebrospin fluid pressure by assessing the extent to which the op disc protrudes into the eyeball.

OBSTRUCTION TO THE FLOW OF CEREBROSPINAL FLUID

Hydrocephalus. "Hydrocephalus" means exce water in the cranial vault. This condition is frequent divided into *communicating hydrocephalus* and *noncom municating hydrocephalus*. In communicating hydr cephalus fluid flows readily from the ventricular syste into the subarachnoid space, while in noncommunica ing hydrocephalus fluid flow out of one or more of t ventricles is blocked.

Usually the *noncommunicating* type of hydrocephal is caused by a block in the aqueduct of Sylvius so th as fluid is formed by the choroid plexuses in the tw lateral and the third ventricles, the volumes of the three ventricles increase greatly. This flattens the bra

into a thin shell against the skull. In newborn babies the increased pressure also causes the whole head to swell because the skull bones still have not fused.

The communicating type of hydrocephalus is usually caused by blockage of fluid flow in the subarachnoid space around the basal regions of the brain or blockage of the arachnoidal villi themselves. Fluid therefore collects both inside the ventricles and on the outside of the brain, causing the head to swell tremendously in infants and usually damaging the brain severely as well.

The most effective therapy for hydrocephalus is surgical institution of a silicone rubber tube shunt all the way from one of the ventricles to the peritoneal cavity, where the fluid can then be absorbed through the peritoneum.

THE BLOOD-CEREBROSPINAL FLUID AND BLOOD-BRAIN BARRIERS

It has already been pointed out that the constituents of the cerebrospinal fluid are not exactly the same as those of the extracellular fluid elsewhere in the body. Furthermore, many large molecular substances hardly pass at all from the blood into the cerebrospinal fluid or into the interstitial fluids of the brain even though these same substances pass readily into the usual interstitial fluids of the body. Therefore, it is said that barriers, called the *blood-cerebrospinal fluid barrier* and the *blood-brain barrier*, exist between the blood and the cerebrospinal fluid and brain fluid, respectively. These barriers exist in the choroid plexus and in essentially all areas of the brain parenchyma *except the hypothalamus*, where substances diffuse with ease into the tissue spaces. This ease of diffusion is very important because the hypothalamus responds to many different changes in the body fluids, such as changes in osmolality, glucose concentration, and so forth; these responses provide the signals for feedback regulation of each of the factors.

In general, the blood-cerebrospinal fluid and blood-brain barriers are highly permeable to water, carbon dioxide, oxygen, and most lipid-soluble substances such as alcohol and most anesthetics; slightly permeable to the electrolytes, such as sodium, chloride, and potassium; and almost totally impermeable to plasma proteins and many large organic molecules. Therefore, the blood-cerebrospinal fluid and blood-brain barriers often make it impossible to achieve effective concentrations of either protein antibodies or some drugs in the cerebrospinal fluid or parenchyma of the brain.

The cause of the low permeability of the blood-cerebrospinal fluid and blood-brain barriers is the manner in which the endothelial cells of the capillaries are joined to each other: They are joined by so-called *tight junctions* That is, the membranes of the adjacent endothelial cells are almost fused with each other rather than having slit-pores between them, as is the case in most other capillaries of the body.

Interstitial Fluid of the Brain. Research studies have shown that the interstitial fluid in brain tissue is about 12 per cent of the tissue weight, in contrast to 17 per cent elsewhere in the body. The cause of this difference is lack of collagen fibers between the cells.

Diffusion Between the Cerebrospinal Fluid and the Brain Interstitial Fluid. The surfaces of the ventricles are lined with a thin cuboidal epithelium called the *ependyma*, and the cerebrospinal fluid on the outer surfaces of the brain is separated from the brain tissue by a thin membrane called the *pia mater*. Both the ependyma and the pia mater are extremely permeable so that almost all substances that enter the cerebrospinal fluid can also diffuse readily into the surface areas of the brain interstitial fluid. Or, likewise, substances in the interstitial fluid can diffuse in the other direction as well. Therefore, many drugs that have no effect at all on the brain when introduced into the blood stream nevertheless can have very important effects on the brain when injected into the cerebrospinal fluid.

Physiological Importance of the Blood-Brain Barrier, and Importance of the Special Composition of the Cerebrospinal Fluid

The neurons of the brain require a very exactly controlled environment, or else their function becomes abnormal and so also does the function of the entire brain The blood-brain barrier protects the cerebral tissue from detrimental substances in the blood, and the transport processes of the choroid plexuses and the brain capillary endothelium help provide the appropriate fluid environment for the brain

One of the most important results of this special brain fluid control system is the low concentration of potassium ions in the brain interstitial fluid. Experiments have shown that even when the circulating blood potassium rises to values almost two times normal, the potassium concentration in the cerebrospinal fluid still remains at its normal low value. Thus, the barrier system, along with its carrier-mediated transport of potassium, not only maintains a low potassium concentration but also keeps this concentration very constant, allowing the neurons to generate very high electrical potentials that do not change with the vagaries of the rest of the body

The blood-brain barrier also prevents such substances as acetylcholine, norepinephrine, dopamine, and glycine from entering the brain from the blood even though their concentrations might become quite high in the circulating blood. This is exceedingly important, because all these are very powerful synaptic transmitter substances and could have devastating effects on brain function.

THE INTRAOCULAR FLUID

The eye is filled with *intraocular fluid*, which maintains sufficient pressure in the eyeball to keep it distended. Figure 32–4 illustrates that this fluid can be divided into two portions, the *aqueous humor*, which lies in front and to the sides of the lens, and the *vitreous humor*, which lies between the lens and the retina. The aqueous humor is a freely flowing fluid, whereas the vitreous humor, sometimes called the *vitreous body*, is a gelatinous mass held together by a fine fibrillar network composed primarily of large proteoglycan molecules. Substances can *diffuse* slowly in the vitreous humor, but there is little *flow* of fluid.

Aqueous humor is continually being formed and reabsorbed. The balance between formation and reabsorption of aqueous humor regulates the total volume and sure of the intraocular fluid.

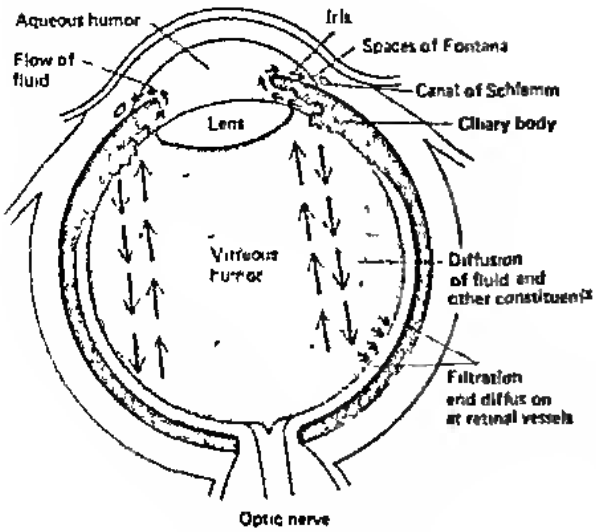


Figure 32–4. Formation and flow of fluid in the eye

FORMATION OF AQUEOUS HUMOR BY THE CILIARY BODY

Aqueous humor is formed in the eye *at an average rate of 1 to 2 cubic millimeters each minute* Essentially all of this is secreted by the *ciliary processes*, which are linear folds projecting from the *ciliary body* into the space behind the iris where the lens ligaments also attach to the eyeball A cross-section of these ciliary processes is illustrated in Figure 32–5, and their relationship to the fluid chambers of the eye can be seen in Figure 32–4 Because of their folded architecture, the total surface area of the ciliary processes is approximately 6 square centimeters in each eye—a large area, considering the small size of the ciliary body The surfaces of these processes are covered by epithelial cells, and immediately beneath these is a highly vascular area

Aqueous humor is formed by the ciliary processes in much the same manner that cerebrospinal fluid is formed by the choroid plexus The postulated mechanism is the following It is believed that the ciliary epithelium actively secretes sodium, chloride, and probably bicarbonate ions into the spaces between the cells This in turn causes osmosis of water into these spaces, and t resulting solution then oozes out of the surfaces of t ciliary processes. In addition, several nutrients a transported across the epithelium by active transport facilitated diffusion; these include amino acids, ascorb acid, and probably also glucose.

OUTFLOW OF AQUEOUS HUMOR FROM THE EYE

After aqueous humor is formed by the ciliary pr esses, it flows, as shown in Figure 32–4, *between t ligaments of the lens*, then *through the pupil*, and final *into the anterior chamber of the eye*. Here, the fluid flo into the *angle between the cornea and the iris* and then through a meshwork of *trabeculae*, finally into the can *of Schlemm* Figure 32–6 illustrates the anatomic structures at the irido-corneal angle, showing that t spaces between the trabeculae extend all the way fro the anterior chamber to the canal of Schlemm. T canal of Schlemm in turn is a thin-walled vein th extends circumferentially all the way around the e Its endothelial membrane is so porous that even lar protein molecules, as well as small particulate matt up to the size of red blood cells, can pass from t anterior chamber into the canal of Schlemm. Ev though the canal of Schlemm is actually a venous blo vessel, so much aqueous humor normally flows into that it is filled only with aqueous humor rather th with blood Also, the small veins that lead from t canal of Schlemm to the larger veins of the eye usual contain only aqueous humor, and these are call *aqueous veins*

INTRAOCULAR PRESSURE

The average normal intraocular pressure is appro mately 16 mm Hg, with a range of from 12 to 20

Tonometry. Because it is impractical to pass a need into a patient's eye for measurement of intraocul pressure, this pressure is measured clinically by mea of a tonometer, the principle of which is illustrated Figure 32–7 The cornea of the eye is anesthetized wi a local anesthetic, and the footplate of the tonometer placed on the cornea A small force is then applied to central plunger, causing the central part of the corne

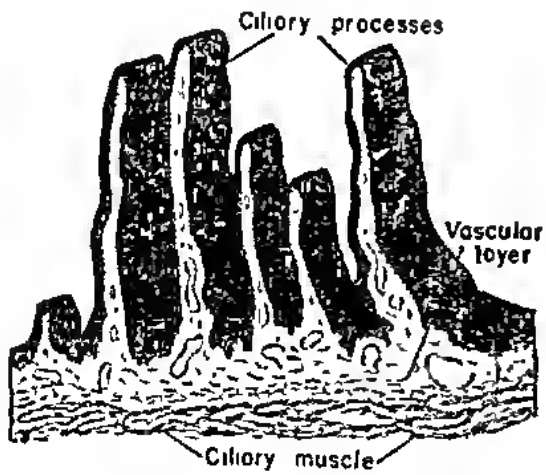


Figure 32–5. Anatomy of the ciliary processes.

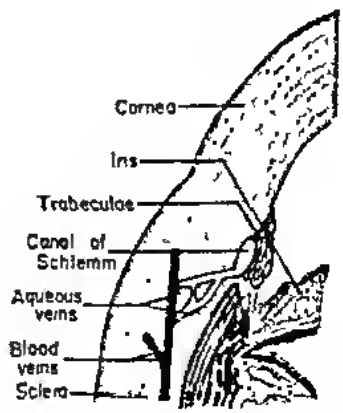


Figure 32–6. Anatomy of the iridocorneal angle, showing th system for outflow of aqueous humor into the conjunctival veins

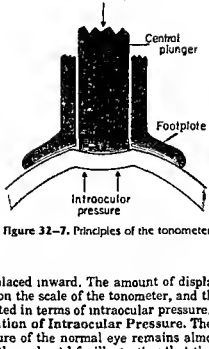


Figure 32–7. Principles of the tonometer

be displaced inward. The amount of displacement is recorded on the scale of the tonometer, and this in turn calibrated in terms of intraocular pressure.

Regulation of Intraocular Pressure. The intraocular pressure of the normal eye remains almost exactly constant throughout life, illustrating that the pressure-regulating mechanism is very effective. But the precise operation of this mechanism is not clear The pressure is regulated mainly by the outflow resistance from the anterior chamber into the canal of Schlemm, presumably in the following way: The wall of the canal of Schlemm consists of roultiple layers, and the intraocular fluid penetratee through the indocorneal angle to the outer surface of this wall. When the pressure of the intraocular fluid is low, the thickness of the wall ie very slight, and the multiple layers remain compacted against each other in such a way that the fluid cannot flow readily into the canal of Schlemm. However, when the intraocular pressure rises higher than about 15 mm Hg, much of the intraocular fluid then penetrates into the wall of the canal of Schlemm, pushing the layers of the cells apart and opening large fluid channels about 2 to 3 microns in size. These, in turn, allow large quantities of intraocular fluid to enter the canal of Schlemm. Thus, this valvelike opening of the spaces through the wall of the canal of Schlemm providee an automatic regulatory system for keeping the intraocular pressure at a nearly constant level day in and day out.

Cleansing of the Trabecular Spacee and of the Intraocular Fluid. When large amounts of debrie occur in the aqueous humor, as occurs following hemorrhage into the eye or during intraocular infection, the debris is likely to accumulate in the trabecular spaces leading to the canal of Schlemm, therefore preventing adequate reabsorption of fluid from the anterior chamber and sometimes causing glaucoma, as explained subsequently. However, on the surfaces of the trabecular plates are large numbers of phagocytic cells Also, immediately outside the canal of Schlemm is a layer of interstitial gel containing large numbers of reticuloendothelial cells that have an extremely high capacity for both engulfing debris and degrading it into small molecular substances that can then be absorbed. Thus, this phagocytic system keeps the trabecular spaces cleaned.

In addition, the surface of the iris and other surfaces of the eye behind the iris are covered with an epithelium that is capable of phagocytizing proteins and small particles from the aqueous humor, thereby helping maintain a perfectly clear fluid.

Glaucoma. Glaucoma is one of the most common causes of blindness. It is a disease of the eye in which the intraocular pressure becomes pathologically high, sometimes rising to as high as 70 mm Hg. Pressures rising above as little as 25 to 30 mm Hg can cause loss of vision when maintained for long periods of time. And the extremely high pressures can cause blindness within days. As the pressure rises, the retinal artery, which enters the eyeball at the optic disc, is compressed, thus reducing the nutrition to the retina. This often results in permanent atrophy of the retina and optic nerve, with consequent blindness.

In essentially all cases of glaucoma the abnormally high pressure results from increased resistance to fluid outflow at the iridocorneal junction In most patients, the cause of this is unknown, but in some it results from infection or trauma to the eye. White blood cells and tissue debris can block the outflow of fluid, thereby greatly increasing the intraocular pressure.

Administration of *Diamox* is often effective in reducing the pressure in glaucoma. This drug acts by reducing the rate of formation of aqueous humor, probably because of an effect that reduces bicarbonata and sodium ion secretion.

Sometimes, when all other therapeutic procedures fail to reinstitute adequate outflow of aqueous humor, a *trephine* operation is performed—that is, a small hole is made through the corneoscleral junction into the anterior chamber, a portion of the iris is removed, and the conjunctiva is sewed over the hole. Fluid thereafter leaks from the anterior chamber into the areolar tissue beneath the conjunctiva and ie absorbed into the lymphatic channels.

FLUID CIRCULATION IN THE POTENTIAL SPACES OF THE BODY

The important potential spates of the body are the *pleural cavity*, the *pericardial cavity*, the *peritoneal cavity*, the cavity of the *tunica vaginalis* surrounding the testis, and the *synovial cavities* that include both the joint cavities and the bursae Normally, these are mainly empty except for a few milliliters of viscous fluid that lubricatea movement of the surfaces against each other. But in certain abnormal conditions the spaces can swell to contain tremendous quantities of fluid, which is the reason they are called "potential" spaces.

FLUID EXCHANGE BETWEEN THE CAPILLARIES AND THE POTENTIAL SPACES

The membrane of a potential space usually does not offer significant resistance to the passage of fluid, electrolytes, or even proteins back and forth between the space and the interstitial fluid of the surrounding tissue. Consequently, fluid leaving a capillary adjacent to the potential space membrane, as shown in Figure 32–8, not only diffuses into the interstitial fluid but also into the potential space. Likewise, fluid can diffuse back out of the space into the interstitial fluid and thence into the capillary.

Lymphatic Drainage of the Potential Spaces. In the same manner that protein collects in the in

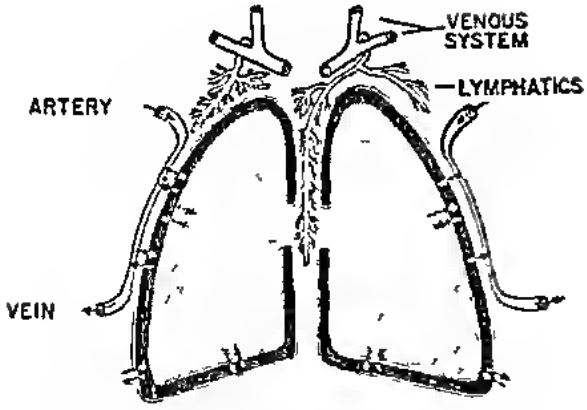


Figure 32–8. Dynamics of fluid exchange in the intrapleural spaces

spaces because of leakage out of the capillaries, so does it also tend to collect in the potential spaces, and it must be removed through lymphatics or some other channel Fortunately, each potential space is either directly or indirectly connected with lymphatic drainage systems, which will be discussed in connection with each different type of potential space.

Function of the Lymphatic Pump in the Drainage of the Potential Spaces. Almost all the potential spaces are located where there is continual movement the joints are moved, the bursae are compressed, the heart moves in the pericardial cavity, the lungs move in the pleural cavity with respiration, and the peritoneal spaces are also squeezed with each respiratory cycle or with other body movement. Each movement compresses fluid into the lymphatics surrounding the potential spaces, and, because of the valves in the lymphatics and the pumping mechanisms of the lymphatics, the fluid flows continually away from the space. Therefore, any time that excess protein collects in the potential space and osmotically pulls any excess fluid whatever into the space, this fluid automatically washes the protein into the lymphatics and from there back into the blood stream.

Altered Capillary Dynamics as a Cause of Increased Pressure and Fluid in the Potential Spaces. Any abnormal changes that can occur in capillary dynamics to cause extracellular edema of the tissue spaces, as described in the preceding chapter, can also cause increased pressure and fluid in the potential spaces Thus, *increased capillary permeability, increased capillary pressure, decreased plasma colloid osmotic pressure,* and *blockage of the lymphatics* from a potential spare can all cause swelling of the space. The fluid that collects is called a *transudate.*

One of the most common causes of swelling in a potential space is *infection.* White blood cells and other debris caused by the infection block the lymphatics, resulting in (1) buildup of protein in the space, (2) increased colloid osmotic pressure, and (3) consequent failure of fluid reabsorption.

CAVITY

Figure 32–8 shows specifically the diffusion of fluid into and out of the pleural cavity at the parietal and the visceral pleural surfaces. This occurs in precisel the same manner as in the usual tissue spaces, exce that a very porous mesenchymal *serous membrane,* t *pleura,* is interspersed between the capillaries and t pleural cavity.

Large numbers of lymphatics drain from the media tinal and lateral surfaces of the parietal pleura, a with each expiration the intrapleural pressure rise forcing small amounts of fluid into the lymphatics, al the respiratory movements alternately compress t lymphatic vessels, promoting continuous flow along t lymphatic channels.

Maintenance of a Negative Pressure in t Pleural Cavity. The visceral pleura of the lungs conti ually absorbs fluid with considerable "absorptive force This is caused by the low capillary pressure—about 5 10 mm Hg—in the pulmonary system. In contrast this low pressure, the plasma proteins exert about 7 mm Hg colloid osmotic pressure, causing an absorpti pressure at the visceral pleura of up to 20 mm Hg at times As a result, the pressure of the *fluid* in t intrapleural space remains negative at all times, ave aging −8 mm Hg This negative pressure is mu greater than the elastic force of the lungs (−4 mm H that tends to collapse the lungs and to pull the lun away from the chest wall. Therefore, it keeps the lun expanded

THE PERICARDIAL CAVITY

The space around the heart operates with essential the same dynamics as those of the pleural cavity. T pressure of the fluid in the pericardial cavity, like th in the pleural cavity, is negative. Here again, durin expiration as well as during excessive filling of t heart, the pericardial pressure rises intermittently, for ing excess fluid into lymphatic channels of the media tinum.

THE PERITONEAL CAVITY

The peritoneal cavity of the abdomen is subject to t same fluid dynamics as are the other potential space fluid is filtered into the peritoneal space through t serous membrane called the *visceral peritoneum* cove ing the viscera and the *parietal peritoneum* lining t outer walls of the abdominal cavity. Fluid is also a sorbed through the peritoneum

The peritoneal cavity is more susceptible to the d velopment of excess fluids than are most of the oth ravities, for two reasons (1) Whenever the pressure the liver sinusoids rises more than 5 to 10 mm Hg abo normal, fluid containing large amounts of protein begi to transude through the liver surface into the abdomin cavity, and (2) the capillary pressure in the viscer peritoneum is higher than elsewhere in the body; th higher pressure is caused by the resistance to port blood flow through the liver. High venous pressu caused by heart failure or extra resistance in the liv in pathological states, such as *cirrhosis, carcinoma,* *portal vein obstruction,* frequently results in marke transudation of fluid into the abdomen; the transuda in this case is called *ascites*

Numerous large lymphatic channels lead from t peritoneal cavity, especially from the lower surface the diaphragm. With each diaphragmatic movemen

relatively large quantities of lymph flow out of the peritoneal cavity into the thoracic duct. This can be shown effectively by injecting radioactive red blood cells into the abdomen. A significant proportion of these cells, still in the whole form, is found in the blood within 10 to 20 minutes. Occasionally, however, cancer spreads so widely throughout the abdomen that it blocks the lymphatics, thereby preventing return of protein to the blood. As a result, the colloid osmotic pressure also rises and severe chronic ascites ensues.

THE SYNOVIAL CAVITIES

The joint cavities and the bursae are known as *synovial cavities*. The synovial membrane is not a true membrane at all but only a collection of dense fibrous tissue cells that line the surface between the interstitial spaces and the cavities. For this reason these cavities might be considered to be nothing more than enlarged tissue spaces. However, the synovial cavities do contain a large amount of proteoglycans, much more than normally present in the interstitial fluids. This is presumably secreted by the surrounding connective tissue cells.

In the synovial cavities, as in the other potential spaces, excess proteins are likely to collect, and these must be returned to the circulatory system through the lymphatics; otherwise the space swells. Since the synovial membrane offers little or no barrier to the transfer of fluid into the surrounding tissues, the protein can flow into the lymphatics of the area.

The pressure in joint cavities usually measures about −4 to −8 mm Hg. This negative pressure results from the same factors that cause negative pressures throughout most of the interstitial spaces of the body; these factors were discussed in detail in the preceding chapter.

REFERENCES

Agostoni, E. Mechanics of the pleural space. *Physiol Rev*, *52*:57, 1972

Allen, L., and Weatherford, T. Role of fenestrated basement membrane in lymphatic absorption from peritoneal cavity. *Am J Physiol*, *197*:551, 1959

Bell, W. E., and McCormick, W. F.: Increased intracranial pressure in children. *Major Probl. Clin Pediatr.*, *8*:3, 1972

Bill, A.: Blood circulation and fluid dynamics in the eye. *Physiol. Rev*, *55*:383, 1975

Bill, A.: Circulation in the eye. *In* Renkin, E. M., and Michel, C. C. (eds.) Handbook of Physiology. Sec. 2, Vol. IV. Bethesda, American Physiological Society, 1984, p. 1001

Chandler, P. A., *et al.* Glaucoma, 2nd Ed. Philadelphia, Lea & Febiger, 1979

Crick, R. P., and Caldwell, A. D. S. (eds.) Glaucoma. New York, Grune & Stratton, 1979

Cserr, H. F.: Physiology of the choroid plexus. *Physiol Rev*, *51*:273, 1971

Davson, H.: The Physiology of the Cerebrospinal Fluid. Boston, Little, Brown, 1967

Davson, H. Physiology of the Eye, 3rd Ed. New York, Churchill Livingstone, Div of Longman, 1972

Elliot, R. H.: A Treatise on Glaucoma. Huntington, N.Y., R. E. Krieger, 1979

Fenstermacher, J. D., and Rapoport, S. I. Blood-brain barrier. *In* Renkin, E. M., and Michel, C. C. (eds.) Handbook of Physiology. Sec. 2, Vol. IV. Bethesda, American Physiological Society, 1984, p. 969

Fischbarg, J., and Lim, J. J. Fluid and electrolyte transports across corneal endothelium. *Curr Top Eye Res*, *4*:201, 1984

Freese, A. S. Glaucoma: Diagnosis, Treatment, Prevention. New York, Public Affairs Committee, 1979

Guyton, A. C., *et al.* Circulatory Physiology II: Dynamics and Control of the Body Fluids. Philadelphia, W. B. Saunders Co., 1975

Hamerman, D., *et al.* The structure and chemistry of the synovial membrane in health and disease. *In* Bittar, E. E., and Bittar, N. (eds.) The Biological Basis of Medicine, Vol. 3. New York, Academic Press, 1969, p. 269

Hills, B. A. The pleural interface. *Thorax*, *40*:1, 1985

Hochwald, G. M. Animal models of hydrocephalus: Recent developments. *Proc Soc Exp Biol Med*, *178*:1, 1985

Katzman, R., and Pappius, H. Brain Electrolytes and Fluid Metabolism. Baltimore, Williams & Wilkins, 1973

Leydhecker, W., and Kroeglstein, G. K. (eds.) Recent Advances in Glaucoma. New York, Springer-Verlag, 1979

Millen, J. W., and Woollam, D. H. M. The Anatomy of the Cerebrospinal Fluid. New York, Oxford University Press, 1962

Notter, R. H., and Finkelstein, J. N. Pulmonary surfactant. An interdisciplinary approach. *J Appl Physiol*, *57*:1613, 1984

Oldendorf, W. H. Blood-brain barrier permeability to drugs. *Annu Rev Pharmacol*, *14*:239, 1974

Pappenheimer, J. R., *et al.* Perfusion of the cerebral ventricular system in unanesthetized goats. *Am J Physiol*, *203*:763, 1962

Saunders, N. R., and Milgard, K. Development of the blood-brain barrier. *J Dev Physiol*, *6*:45, 1984

Shulman, K. (ed.) Intracranial Pressure IV. New York, Springer-Verlag, 1980

Siesjo, B. K. Cerebral circulation and metabolism. *J Neurosurg*, *60*:883, 1984

33

Partition of the Body Fluids: Osmotic Equilibria Between Extracellular and Intracellular Fluids

The body fluids are so important to the basic physiology of bodily function that we discuss them at several points in this text but each time in different contexts In Chapter 9 it was pointed out that the body fluids can be divided mainly into *extracellular* and *intracellular fluids*, and the basic differences between these were discussed, as well as the manner in which these differences come about as the result of cell membrane transport. In Chapter 22 the relationship of blood volume to the overall regulation of the circulation was presented, and in Chapters 30 through 32 the interrelationships between capillary dynamics and interstitial fluids were discussed.

In the present chapter and in the following chapters on the kidneys, we will be concerned with the body fluids from the total point of view, including regulation of body fluid volume, regulation of the constituents in the extracellular fluid, regulation of acid-base balance, and factors that govern gross interchange of fluid between the extracellular and the intracellular compartments—especially the osmotic relationships between these compartments

TOTAL BODY WATER

The total amount of water in a man of average weight (70 kg) is approximately 40 liters (see Fig 33-1), averaging 57 per cent of his total body weight. In a newborn infant this may be as high as 75 per cent of the body weight, but it progressively decreases from birth to old age, most of the decrease occurring in the first 10 years of life Also, obesity decreases the percentage of water in the body, sometimes to as low as 45 per cent

• VERSUS OUTPUT OF WATER

Daily Intake of Water. Most of our daily intake of water enters by the oral route. Approximately two thirds is in the form of pure water or some other beverage, an the remainder is in the food that is eaten A smal amount is also synthesized in the body as the result oxidation of hydrogen in the food; this quantity range between 150 and 250 ml per day, depending on the rat of metabolism The normal intake of fluid, includin that synthesized in the body, averages about 2300 m per day.

Daily Loss of Body Water. Table 33–1 shows th routes by which water is lost from the body unde different conditions Normally, at an atmospheric tem perature of about 68 ° F, approximately 1400 ml of th 2300 ml of water intake is lost in the *urine*, 100 ml lost in the *sweat*, and 100 ml in the *feces*. The remainin 700 ml is lost by *evaporation from the respiratory tra* or by *diffusion through the skin*

Insensible Water Loss. Loss of water by diffusion through the skin and by evaporation from the respira tory tract is known as *insensible water loss* because w

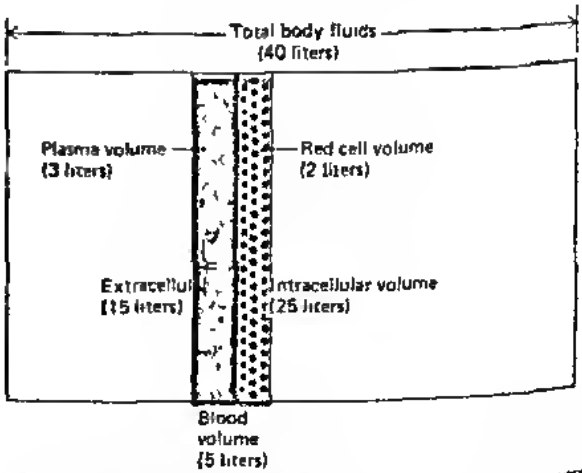


Figure 33–1. The body fluids diagrammed, showing the extra cellular fluid volume, intracellular fluid volume, blood volume, an total body fluids.

Table 33–1 DAILY LOSS OF WATER (In Milliliters)

	Normal Temperature	Hot Weather	Prolonged Heavy Exercise
Insensible loss.			
Skin	350	350	350
Respiratory tract	350	250	650
Urine	1400	1200	500
Sweat	100	1400	5000
Feces	100	100	100
TOTAL	2300	3300	6600

do not know that we are actually losing water at the time that it is leaving the body.

The average loss of water by diffusion through the skin is approximately 300 to 400 ml per day; this amount is lost even in a person who is born without sweat glands. In other words, the water molecules themselves actually diffuse through the cells of the skin. Fortunately, the cholesterol-filled, cornified layer of the skin acts as a protector against still much greater loss of water by diffusion. When the cornified layer becomes denuded, as after extensive burns, the rate of evaporation can increase to as much as 3 to 5 liters each day.

All air that enters the respiratory tract becomes totally saturated with moisture, to a vapor pressure of approximately 47 mm Hg, before it is expelled. Since the vapor pressure of the inspired atmospheric air is usually far less than 47 mm Hg, the average water loss through the lungs is about 300 to 400 ml per day. Because the atmospheric vapor pressure normally decreases with decreasing temperature, the loss is greatest in very cold and least in very warm weather. This explains the dry feeling in the respiratory passages in cold weather.

Loss of Water in Hot Weather and During Exercise. In very hot weather, water loss in sweat is occasionally increased to as much as 1.5 to 2.0 liters an hour, which obviously can rapidly deplete the body fluids. Sweating will be discussed in Chapter 72.

Exercise increases the loss of water in two ways. First, it increases the rate of respiration, which promotes increased water loss through the respiratory tract in proportion to the increased ventilatory rate. Second, and much more important, exercise increases the body heat and consequently is likely to result in excessive sweating.

BODY FLUID COMPARTMENTS

THE INTRACELLULAR FLUID COMPARTMENT

About 25 of the 40 liters of fluid in the body are inside the approximately 75 trillion cells of the body and are collectively called the *intracellular fluid*. The fluid of each cell contains its own individual mixture of different constituents, but the concentrations of these constituents are reasonably similar from one cell to another. For this reason the intracellular fluid of all the different cells is considered to be one large fluid compartment, though in reality it is an aggregate of trillions of minute compartments.

THE EXTRACELLULAR FLUID COMPARTMENT

All the fluids outside the cells are called *extracellular fluid*, and these fluids are constantly mixing, as was explained in Chapter 1. The total amount of fluid in the extracellular compartment averages 15 liters in a 70 kg adult.

The extracellular fluid can be divided into *interstitial fluid, plasma, cerebrospinal fluid, intraocular fluid, fluids of the gastrointestinal tract,* and *fluids of the potential spaces.*

Interstitial Fluid. The interstitial fluid lies in the spaces between the cells. A minute portion of it is free in the form of actual flowing fluid, while probably more than 99 per cent of it is held in the gel of the interstitial spaces, as discussed in Chapter 30.

The fluids in the cerebrospinal system, the eye, the potential spaces, and the gastrointestinal tract are discussed in Chapter 32 and 64. These fluids, on the average, are similar to interstitial fluid and often are considered to be part of the interstitial compartment.

Plasma. The plasma is the noncellular portion of blood. It is part of the extracellular fluid and communicates continually with the interstitial fluid through pores in the capillaries. Loss of plasma from the circulatory system through the capillary pores is minimized by the colloid osmotic pressure exerted by the plasma proteins, which was explained in Chapter 30. Yet the capillaries are porous enough for most dissolved substances and water molecules to *diffuse* through them freely, allowing constant mixing between the plasma and interstitial fluid of almost all substances except the protein.

The plasma volume averages 3 liters in this normal adult.

BLOOD VOLUME

Blood contains both extracellular fluid (the fluid of the plasma) and intracellular fluid (the fluid in the red blood cells). However, since blood is contained in a closed chamber all its own—the circulatory system—its volume and its special dynamics are exceedingly important.

The average blood volume of a normal adult is almost exactly 5000 ml. On the average, approximately *3000 ml of this is plasma,* and the remainder, *2000 ml is red blood cells.* However, these values vary greatly in different individuals; also, sex, weight, and many other factors affect the blood volume.

Effect of Weight and Sex on Blood Volume. In persons who have a minimum of adipose tissue the blood volume varies almost directly in proportion to the body weight, normally averaging about 79 ml/kg ± 10 per cent for both *lean* males and *lean* females. However, the greater the ratio of fat to body wei[illegible] volume per unit weight, because fat [illegible] vascular volume. Figure 33–2 illustrates [illegible] fat, showing that the heavier the person

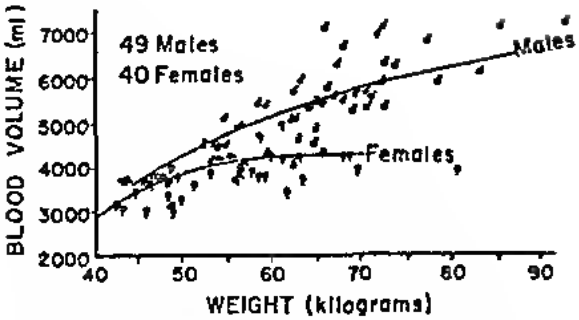


Figure 33–2. Relationship of blood volume to gender and weight

less on the average becomes the blood volume in relation to weight. Furthermore, in very heavy females, who are especially prone to have high ratios of fat tissue to lean tissue, the blood volume per kilogram decreases much more than for males

The Hematocrit. The hematocrit is the percentage of red blood cells in the blood as determined by centrifuging blood in a "hematocrit tube" until the cells become packed tightly in the bottom of the tube. The percentage of red blood cells in the blood can be determined roughly from the level of the packed cells. Such centrifuged blood in hematocrit tubes was illustrated in Figure 18–2 of Chapter 18. Unfortunately, it is impossible for the red blood cells to be packed completely together; about 3 to 8 per cent of plasma remains entrapped among the cells. Therefore, the true hematocrit (H) averages about 96 per cent of the measured hematocrit (Hct), that is:

$$H = 0.96 \text{ Hct}$$

The true hematocrit (H) is approximately 40 for a normal man and 36 for a normal woman

In severe anemia, the hematocrit may fall to as low as 10, but this small quantity of red blood cells is barely sufficient to sustain life On the other hand, a few conditions cause excessive production of red blood cells, resulting in polycythemia. In these instances, the hematocrit often rises to 65, and occasionally to 80. Obviously, there is an upper limit to the level of the hematocrit in polycythemic blood, because excessive hematocrit causes the blood to become so viscous that death ensues as a result of multiple plugging of the peripheral vascular tree.

The Body Hematocrit. The hematocrit of the blood is *considerably less in the capillaries, arterioles, and other very small vessels* than in the large veins and arteries. The cause of this is *axial streaming* of blood cells in blood vessels, which was explained in Chapter 18. In general, red cells cannot flow near the walls of the vessels nearly so easily as can plasma. Therefore, the cells tend to [illegible] the center of the vessels, while a large [illegible]ion of the plasma remains near the walls. In [illegible] very large vessels the ratio of wall surface to total volume is slight, so that the accumulation of plasma near the walls does not affect the hematocrit significantly. However, in the small vessels this ratio of wall surface to volume is great, causing the ratio of plasma to cells to be far greater than in the large vessels

If one averages the hematocrit in both the large and small vessels, a value called *body hematocrit* is determined. For normal man the body hematocrit (H_0) averages 91 per cent of the large vessel hematocrit; that is:

$$H_0 = 0.91 \text{ H} = 0.87 \text{ Hct}$$

MEASUREMENT OF BODY FLUID VOLUMES

THE DILUTION PRINCIPLE FOR MEASURING FLUID VOLUMES

The volume of any fluid compartment of the body can be measured by placing a substance in the compartment, allowing it to disperse evenly throughout the fluid, and then measuring the extent to which the substance has become diluted. Figure 33–3 illustrates this "dilution" principle for measuring the volume of any fluid compartment of the body. In this example, a small quantity of dye or other foreign substance is placed in fluid chamber A, and the substance is allowed to disperse throughout the chamber until it becomes mixed in equal concentrations in all areas, as shown in chamber B. Then a sample of the dispersed fluid is removed and the concentration of the substance is analyzed chemically, photoelectricolly, or by any other means. The volume of the chamber can then be determined from the following formula.

$$\text{Volume in ml} = \frac{\text{Quantity of test substance instilled}}{\text{Concentration per ml of dispersed fluid}}$$

Note that all one needs to know is (1) the *total quantity of the test substance* put into the chamber,

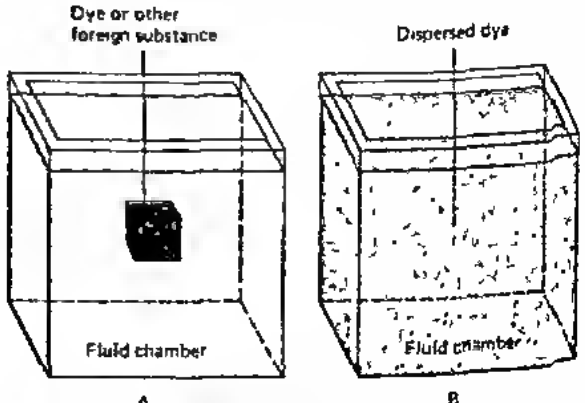


Figure 33–3. Principles of the dilution method for measuring fluid volumes (explained in the text).

and (2) the *concentration in the fluid after dispersement.* As an example, if 10 milligrams of dye is dispersed in a chamber and the final concentration is 0.01 milligram for ml of fluid, then the volume calculates to be 10 divided by 0.01, or 1000 ml.

DETERMINATION OF BLOOD VOLUME

Substances Used in Determining Blood Volume. A substance used for measuring blood volume must be capable of dispersing throughout the blood with ease, and it must remain in the circulatory system long enough for measurements to be made. The two major groups of substances that satisfy these conditions are (1) substances that combine with the red blood cells and (2) substances that combine with the plasma proteins, for both the red blood cells and the plasma proteins remain reasonably well in the circulatory system, and any foreign substance that combines with either of them likewise remains in the blood stream.

Examples of substances that combine with red blood cells and that are used for determining blood volume are *radioactive iron, radioactive chromium,* and *radioactive phosphate.* Examples of substances that combine with plasma proteins are the *vital* dyes and *radioactive iodine.*

Radioactive Red Blood Cells. The method most often used to make red blood cells radioactive is to tag the red blood cells with radioactive chromium (Cr^{51}). A small quantity of Cr^{51} is mixed with a few milliliters of blood previously removed from the person, and this is incubated at 36 ° C for half an hour or more. After this time, most of the Cr^{51} will have entered the red blood cells, but to remove the extra chromium from the mixture, the red blood cells are washed in saline. Their total content of Cr^{51} is then determined with a scintillation counter (apparatus for measuring the total number of radioactive disintegrations occurring in the sample per minute). Then the radioactive cells are reinjected into the person. After mixing in the circulatory system has continued for approximately ten minutes, blood is removed from the circulatory system, and the radioactivity in this blood is determined. Using the above dilution formula, the total blood volume is calculated.

To be accurate, this calculated blood volume must now be corrected to determine the true value, because the blood sample is removed from the veins where the hematocrit is not equal to the body hematocrit, as discussed earlier in the chapter. This correction is made as follows:

$$\text{Actual blood volume} = 1.1 \times \text{Measured blood volume}$$

Vital Dyes for Measurement of Plasma Volume. A number of dyes, generally known as "vital dyes," have the ability to combine with proteins. When such a dye is injected into the blood, it immediately forms a tight union with the plasma proteins. Thereafter, the dye travels where the proteins travel.

The dye almost universally used for measuring plasma volume is *T-1824,* also called *Evans blue.* In making determinations of plasma volume, a known quantity of the dye is injected, and it immediately combines with the proteins and disperses throughout the circulatory system within approximately ten minutes. A sample of the blood is then taken, and the red blood cells are removed from the plasma by centrifugation. Then, by spectrophotometric analysis of the plasma, one can determine the exact quantity of dye in the sample of plasma. From the determined quantity of dye in each milliliter of plasma and the known quantity of dye injected, the *plasma volume* is calculated using the above dilution formula.

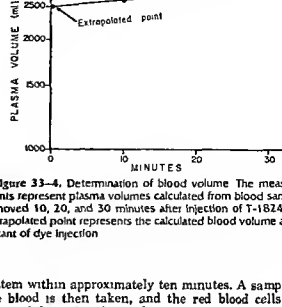


Figure 33–4. Determination of blood volume. The measured points represent plasma volumes calculated from blood samples removed 10, 20, and 30 minutes after injection of T-1824. The extrapolated point represents the calculated blood volume at the instant of dye injection.

To be even more exact in measuring the plasma volume, the rate of loss of dye from the circulatory system during the interval of mixing must also be considered. On the average, T-1824 is lost from the circulatory system at a rate of about 5 per cent per hour, part of this being excreted into the urine and part being carried into the interstitial spaces by the leakage of plasma proteins through the capillary walls. To offset this error, three different samples of blood are usually removed from the circulation at 10, 20, and 30 minutes. Then the plasma volumes are calculated from these samples and plotted on semilog graph paper, as illustrated in Figure 33–4. A straight line is drawn through the measured points and extrapolated back to zero time. Since the percentage loss is approximately the same during the first ten minutes as during each of the two succeeding ten-minute intervals, the extrapolated point indicates approximately the true plasma volume. In Figure 33–4 this value is 2500 ml.

Note that neither T-1824 nor any other vital dye enters the red blood cells. Therefore, this method *does not measure the total blood volume.* However, the blood volume can be calculated from the plasma volume, provided the hematocrit is determined, by using the following formula:

$$\text{Blood volume} = \text{Plasma volume} \times \frac{100}{100 - 0.87\ \text{Hematocrit}}$$

Radioactive Protein. If a sample of plasma is allowed to incubate with radioactive iodine (I^{131}) for 30 minutes or more, some of the protein combines with the iodine, and the iodinated protein can be separated from the remaining iodine by dialysis. The radioactive protein is then injected into the subject, and plasma and blood volumes are determined in the same manner as that discussed for the vital dyes.

MEASUREMENT OF THE EXTRACELLULAR FLUID VOLUME

To use the dilution principle for measuring the volume of the extracellular fluid, one injects into the blood stream a substance that can diffuse readily throughout the entire extracellular fluid chamber, passing easily through the capillary membranes but as little as possible through the cell membranes into the cells After half an hour or more of mixing, a sample of extracellular fluid is obtained by removing blood and separating the plasma from the cells by centrifugation The plasma, which is actually a part of the extracellular fluid, is then analyzed for the injected substance

Substances Used in Measuring Extracellular Fluid Volume—The Concept of "Fluid Space." Substances that have been used for measuring extracellular fluid volume are *radioactive sodium, radioactive chloride, radioactive bromide, radiotiothalomate, thiosulfate ion, thiocyanate ion, inulin,* and *sucrose* Some of these, inulin especially, do not diffuse readily into all the out-of-the-way places of the extracellular fluid compartment Therefore, the volume of extracellular fluid measured with inulin is likely to be lower than the actual volume of the compartment On the other hand, others of these substances—radioactive chloride, radioactive bromide, radioactive sodium, and thiocyanate ion, for instance—are likely to penetrate into the cells to a slight extent and, therefore, are likely to measure a volume somewhat in excess of the extracellular fluid volume

Because no single substance measures the exact extracellular fluid volume, one usually speaks of the *sodium space*, the *thiocyanate space*, the *inulin space*, and so forth, rather than the extracellular volume At present it is impossible to say exactly which "space" most nearly approximates the true extracellular volume. Measurements for the normal 70 kg adult, when different substances have been used, have ranged from 9 liters to 22 liters; *the average measurement has been about 15 liters*

MEASUREMENT OF TOTAL BODY WATER

The total body water can be measured in exactly the same way as the extracellular fluid volume except that a substance must be used that will diffuse into the cells as well as throughout the extracellular compartment The substances that give best results are either *tritiated water* (which can be analyzed with radiation-measuring instruments) or *heavy water* (which can be analyzed quantitatively either by accurate specific gravity measurements of water samples or by infrared spectrophotometry) After administration of the "tagged" water, several hours are required for complete mixing with all the water of the body, and appropriate corrections must be made for any fluid that is lost either into the urine or otherwise during this period of mixing The concentration of the tagged water in the total body water can, at the end of the period of mixing, be determined by simply measuring the tagged water concentration in the plasma.

Measurements of total body water in the 70 kg adult ranged from as low as 30 liters to as high as 50 liters, with a *reasonable average of approximately 40 liters*, or 57 per cent of the total body mass.

CALCULATION OF INTERSTITIAL FLUID VOLUME

Since any substance that passes into the interstitial fluid also passes into almost all other portions of the extracellular fluid, there is no direct method for measuring interstitial fluid volume separately from the entire extracellular fluid volume. However, if the extracellular fluid volume and plasma volume have both been measured, the interstitial fluid volume can be approximated by *subtracting the plasma volume from the total extracellular fluid volume.* This calculation gives *a normal interstitial fluid volume of 12 liters* in the 70 kg adult.

CONSTITUENTS OF EXTRACELLULAR AND INTRACELLULAR FLUIDS

Figure 33–5 diagrams the major constituents of the extracellular and intracellular fluids. (The actual values for these constituents were given in tabular form in Figure 9–1 of Chapter 9.) The quantities of the different substances are represented in Figure 33–5 in *milliequivalents* or *millimoles per liter.* However, the protein molecules and some of the nonelectrolyte molecules are extremely large compared with the more numerous small ions. Therefore, *in terms of mass,* the proteins and nonelectrolytes actually comprise about 90 per cent of the dissolved constituents in the plasma, about 60 per cent in the interstitial fluid, and about 97 per cent in the intracellular fluid.

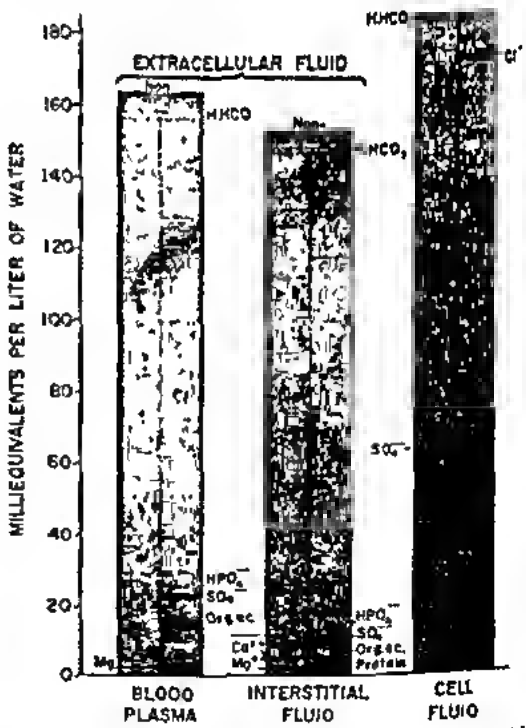


Figure 33–5. The compositions of plasma, interstitial fluid, and intracellular fluid. (Modified from Gamble: Chemical Anatomy Physiology, and Pathology of Extracellular Fluid A Lecture Syllabus Harvard University Press 1954.)

Figure 33–6 illustrates the distribution of the nonelectrolytes in the plasma, and most of these same substances are also present in almost equal concentrations in the interstitial fluid, except for some of the fatty compounds that are present in the plasma in large suspended particles, the *lipoproteins*.

Important Constituents of the Extracellular Fluid. Referring again to Figure 33–5, one sees that extracellular fluid, both that of the blood plasma and of the interstitial fluid, contains large quantities of sodium and chloride ions, reasonably large quantities of *bicarbonate ion*, but only small quantities of potassium, calcium, magnesium, phosphate, sulfate, and organic acid ions. In addition, plasma contains a large amount of protein whereas interstitial fluid contains much less. (The proteins in these fluids and their significance were discussed in Chapter 30.)

In Chapter 1 it was pointed out that the extracellular fluid is called the *internal environment* of the body and that its constituents are accurately regulated so that the cells remain bathed continually in a fluid containing the proper electrolytes and nutrients for continued cellular function. The regulation of most of these constituents will be presented in Chapter 36.

Important Constituents of the Intracellular Fluid. From Figure 33–5 it is also clear that the intracellular fluid contains only small quantities of sodium and chloride ions and almost no calcium ions; but it does contain large quantities of *potassium* and *phosphate* and moderate quantities of *magnesium* and *sulfate ions*, all of which are present in only small concentrations in the extracellular fluid. In addition, cells contain very large concentrations of protein, approximately four times as large as in plasma.

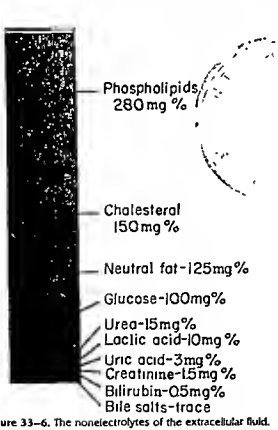


Figure 33–6. The nonelectrolytes of the extracellular fluid.

OSMOTIC EQUILIBRIA AND FLUID SHIFTS BETWEEN THE EXTRACELLULAR AND INTRACELLULAR FLUIDS

One of the most troublesome of all problems in clinical medicine is maintenance of adequate body fluids and proper balance between the extracellular and intracellular fluid volumes in seriously ill patients. The purpose of the following discussion, therefore, is to explain the interrelationships between extracellular and intracellular fluid volumes and the osmotic factors that cause shifts of fluid between the extracellular and intracellular compartments.

BASIC PRINCIPLES OF OSMOSIS AND OSMOTIC PRESSURE

The basic principles of osmosis and osmotic pressure were presented in Chapter 9. However, these principles are so important to the following discussion that they are reviewed here briefly.

Whenever a membrane between two fluid compartments is permeable to water but not to some of the dissolved solutes—that is, the membrane is selectively *permeable*—and the concentration of the solutes is greater on one side of the membrane than on the other, water passes through the membrane toward the side with the greater concentration of solutes. This phenomenon is called *osmosis*.

Osmosis results from the kinetic motion of the individual particles—both the molecules and the ions—in the solutions on the two sides of the membrane. This can be explained in the following way: When the temperature is the same in both solutions, the particles on both sides of the membrane, on the average, all have the same amount of kinetic energy of motion. However, the nondiffusible particles in the two solutions displace some of the water molecules. As a result the so-called *chemical potential* of the water is reduced on each side of the membrane in direct proportion to the concentration of nondiffusible particles. Therefore, on the side of the membrane where there is the lower concentration of nondiffusible particles, the concentration of water molecules is the greater; that is, the chemical potential of the water is greater. This means that more water molecules will strike each pore of the membrane each second on the side of the membrane that has the lowest concentration of nondiffusible solutes. As a result, net diffusion of water molecules will occur from this side toward the solution on the opposite side

of the membrane where the concentration of nondiffusible particles is greater. This net rate of diffusion is called the *rate of osmosis.*

Osmotic Pressure. Osmosis of water molecules can be opposed by applying a pressure across the selectively permeable membrane in the direction opposite that of the osmosis. The amount of pressure required to oppose the osmosis exactly is called the *osmotic pressure.*

Relationship of the Molecular Concentration of a Solution to Its Osmotic Pressure. Each nondiffusible molecule dissolved in water changes the chemical potential of the water by a given amount. Consequently, the tendency for this to cause osmosis of water is directly proportional to the concentration of the nondiffusible molecules. And, as a corollary, the osmotic pressure of the solution is also proportional to the concentration of the nondiffusible molecules in the solution. This relationship holds true for all nondiffusible molecules almost regardless of their molecular weight. For instance, one molecule of albumin with a molecular weight of 70,000 has the same osmotic effect as a molecule of glucose with a molecular weight of 180.

Osmotic Effect of Ions. Nondiffusible ions cause osmosis and osmotic pressure in exactly the same manner as do nondiffusible molecules. Furthermore, when a molecule dissociates into two or more ions, each of the ions then exerts osmotic pressure individually. Therefore, to determine the osmotic effect, all the nondiffusible ions must be added to all the nondiffusible molecules; but note that a bivalent ion, such as calcium, exerts no more osmotic pressure than does a univalent ion, such as sodium.

Osmoles. The ability of solutes to cause osmosis and osmotic pressure is measured in terms of "osmoles"; the osmole is a measure of the total number of particles. *One gram mole of nondiffusible and nonionizable substance is equal to 1 osmole.* On the other hand, if a substance ionizes into two ions (sodium chloride into sodium and chloride ions, for instance), 0.5 gram mole of the substance equals 1 osmole. The obvious reason for using the osmole is that osmotic pressure is determined by the number of particles instead of the mass of the solute.

In general, the osmole is too large a unit for satisfactory use in expressing osmotic activity of solutes in the body. Therefore, the term *milliosmole,* which equals $1/1000$ osmole, is commonly used.

Osmolality and Osmolarity. The osmolal concentration of a solution is called its *osmolality* when the concentration is expressed in osmoles per kilogram of water; it is called *osmolarity* when it is expressed as osmoles per liter of solution. The term "osmolality" is generally preferred because the osmotic pressure of a solution is considerably more closely related to its osmolality than to its osmolarity in very concentrated solutions. However, in the very dilute solutions of the normal human body, the differences are so slight that the terms are frequently used interchangeably. Furthermore, it is so much easier to express the body fluid quantities in liters than in kilograms of water that almost all calculations are based on osmolarities rather than on osmolalities. It is common practice for many physiologists to speak in terms of osmolality even though they make their calculations in terms of osmolarities.

Relationship of Osmotic Pressure to Osmolality. The osmotic pressure of a solution *at body temperature* can be determined approximately from the following formula:

$$\text{Osmotic pressure (mm Hg)} = 19.3 \times \text{Osmolality (milliosmole/liter)}$$

OSMOLALITY OF THE BODY FLUIDS

Table 33–2 lists the osmotically active substances in plasma, interstitial fluid, and intracellular fluid. The milliosmoles of each of these per liter of water is given. Note especially that approximately four fifths of the total osmolality of the interstitial fluid and plasma is caused by sodium and chloride ions, whereas approximately half of the intracellular osmolality is caused by potassium ions; the remainder is divided among the many other intracellular substances.

As noted at the bottom of Table 33–2, the total osmolality of each of the three compartments is approximately 300 milliosmoles per liter, with that of the plasma 1.3 milliosmoles greater than that of the interstitial and intracellular fluids. This slight difference between plasma and interstitial fluid is caused by the osmotic effect of the plasma proteins, which maintains about 23 mm Hg greater pressure in the capillaries than in the surrounding interstitial fluid spaces, as was explained in Chapter 30.

Corrected Osmolar Activity of the Body Fluids. At the bottom of Table 33–2 is shown a *corrected osmolar activity* of plasma, interstitial fluid, and intracellular fluid. The reason for this correction is the following: All molecules and ions in solution exert either *intermolecular attraction* from one solute molecule to the next or *intermolecular repulsion,* and these two effects can cause, respectively, a decrease or an increase in the osmotic "activity" of the dissolved substance. In general, there is more intermolecular attraction than repulsion, so that the overall osmotic activity of the substances is only about 93 per cent of that which one would calculate from the number of milliosmoles present. For this reason, the actual osmotic pressure of the body fluids is proportional to the corrected osmolar activity, which amounts to approximately 280 milliosmoles/liter.

However, since osmotic effects in the body are

Table 33–2 OSMOLAR SUBSTANCES IN EXTRACELLULAR AND INTRACELLULAR FLUIDS

	Plasma (mOsmole/L of H_2O)	Interstitial (mOsmole/L of H_2O)	Intracellular (mOsmole/L of H_2O)
Na^+	146	142	14
K	4 2	4 0	140
Ca^{++}	2 5	2 4	0
Mg^{++}	1 5	1 4	31
Cl	105	108	4
HCO_3	27	28 3	10
HPO_4 . H_2PO_4	2	2	1 1
SO_4	0 5	0 5	1
Phosphocreatine			45
Carnosine			14
Amino acids	2	2	8
Creatine	0 2	0 2	9
Lactate	1 2	1 2	1 5
Adenosine triphosphate			5
Hexose monophosphate			3 7
Glucose	5 6	5 6	
Protein	1 2	0 2	4
Urea	4	4	4
TOTAL mOsmole/L	302 9	301 8	302.2
Corrected osmolar activity (mOsmole/L)	282 6	281 3	281 3
Total osmotic pressure at 37°C (mm hg)	5453	5430	5430

usually determined by relative, rather than absolute, osmolar concentrations, the correction factor is usually ignored.

Total Osmotic Pressure Exerted by the Body Fluids. At the bottom of Table 33–2 is shown the total *osmotic pressure* in mm Hg that would be exerted by each of the different fluids if it were placed on one side of a cell membrane with pure water on the other side Note that this total pressure averages about 5450 mm Hg for plasma, which is 19.3 times the corrected osmolality of 282 6 milliosmoles per kilogram for plasma. Also note that the osmotic pressure of plasma is 23 mm Hg greater than that of the interstitial fluids, this difference equaling the approximate hydrostatic pressure difference between the pressure of the blood inside the capillaries and the pressure in the interstitial fluid outside the capillaries.

MAINTENANCE OF OSMOTIC EQUILIBRIUM BETWEEN EXTRACELLULAR AND INTRACELLULAR FLUIDS

The tremendous osmotic pressure that can develop across the cell membrane when one side of the membrane is exposed to pure water—more than 5400 mm Hg—illustrates how much force can become available to move water molecules through the membrane when the solutions on the two sides of the membrane are not in osmotic equilibrium. For instance, in Figure 33–7A, a cell is placed in a solution that has an osmolality far less than that of the intracellular fluid. As a result, osmosis of water into the cell begins immediately, causing the cell to swell and diluting the intracellular fluid while concentrating the extracelluler fluid. When the fluid inside the cell becomes diluted sufficiently to equal the osmolal concentration of the fluid on the outside, as shown in Figure 33–7B, further osmosis then ceases.

In Figure 33–7C, a cell is placed in a solution having a much higher concentration outside the cell than inside. This time, water passes by osmosis to the exterior, concentrating the intracellular fluid and diluting the extracellular fluid. In this process the cell shrinks until the two concentrations become equal, as shown in Figure 33–7D.

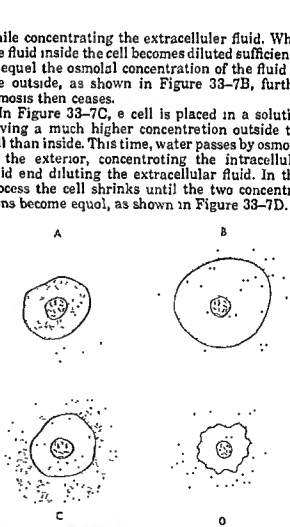


Figure 33–7. Establishment of osmotic equilibrium when cells are placed in a hypo- or hypertonic solution.

Rapidity of Attaining Extracellular and Intracellular Osmotic Equilibrium. The transfer of water through the cell membrane by osmosis occurs so rapidly that any lack of osmotic equilibrium between the two fluid compartments in any given tissue is usually corrected within a few seconds and at most within a minute or so. However, this rapid transfer of water does not mean that complete equilibration occurs between the extracellular and intracellular compartments throughout the whole body within this same short period of time. The reason for this is that fluid usually enters the body through the gut and must then be transported by the blood to all tissues before complete equilibration can occur. In the normal person it may take as long as 30 minutes to achieve reasonably good equilibration everywhere in the body after drinking water.

Isotonicity, Hypotonicity, and Hypertonicity. A fluid into which normal body cells can be placed without causing either swelling or shrinkage of the cells is said to be *isotonic* with the cells. A 0 9 per cent solution of sodium chloride or a 5 per cent glucose solution is approximately isotonic Isotonic solutions are especially important in clinical medicine because they can be infused into the blood without danger of upsetting the osmotic equilibria between the fluids outside and inside the cells.

A solution that will cause the cells to swell is said to be *hypotonic;* any solution of sodium chloride with less than 0.9 per cent concentration is hypotonic.

A solution that will cause the cells to shrink is said to be *hypertonic,* sodium chloride solutions of greater than 0.9 per cent concentration are all hypertonic

CHANGES IN THE VOLUMES AND OSMOLALITIES OF THE EXTRACELLULAR AND INTRACELLULAR FLUID COMPARTMENTS IN ABNORMAL STATES

CALCULATION OF FLUID SHIFTS BETWEEN THE EXTRACELLULAR AND INTRACELLULAR FLUID COMPARTMENTS

Among the different factors that can cause extracellular or intracellular volumes to change markedly are ingestion of water, dehydration, intravenous infusion of different types of solutions, loss of large quantities of fluids from the gastrointestinal tract, or loss of abnormal quantities of fluid by sweating or through the kidneys

The changes in both extracellular and intracellular fluid volumes can be calculated easily if one will keep the following two basic principles in mind

(1) The osmolalities of the extracellular and intracellular fluids remain exactly equal to each other except for a few minutes after a change in one of the fluids occurs

(2) The number of osmoles of osmotically active substance in each compartment, in the extracellular fluid or in the intracellular fluid, remains constant unless one of the osmotically active substances moves through the cell membranes to the other compartment.

Using these two basic principles, we can now analyze the effects of different abnormal fluid conditions on extracellular and intracellular fluid volumes and osmolalities

EFFECT OF ADDING WATER TO THE EXTRACELLULAR FLUID

Water can be added to the extracellular fluid by injection into the blood stream, by injection beneath the skin, or by ingesting water followed by absorption from the gastrointestinal tract into the blood. The water dilutes the extracellular fluid, causing it to become hypotonic with respect to the intracellular fluids Osmosis begins immediately at the cell membranes, with large amounts of water passing to the interiors of the cells Within a few minutes the water becomes distributed almost evenly among all the extracellular and intracellular fluid compartments.

Calculation of the Changes in Extracellular and Intracellular Volumes and Osmolalities. Table 33-3 illustrates the progressive changes that would result from injecting 10 liters of water into the extracellular fluid. This table gives the volumes and milliosmoles per liter of the extracellular fluid, the intracellular fluid, and the total body water

Initially the extracellular fluid volume is 15 liters, the intracellular fluid volume 25 liters, and the total body water 40 liters, the milliosmoles per liter is 300 in each of these. The solution that is added, 10 liters of water, has zero osmolality, and when first injected it is added both to the extracellular fluid and to the total body water, but not to the intracellular fluid.

The third line of the table shows the instantaneous effect on the volumes and osmolalities caused by addition of this fluid (before any osmosis occurs). The extracellular fluid volume rises to 25 liters, and its concentration becomes diluted immediately to 180 milliosmoles/liter. During this instantaneous interval of time nothing has happened to the intracellular fluids. But within a few seconds to a few minutes, the dilute extracellular fluid diffuses throughout the interstitial spaces of the body and comes in contact with the cells, allowing large portions of the water to pass by osmosis into the intracellular fluid Osmosis will not stop until the milliosmolalities of the two fluids become equal. As a result, the milliosmolality of the extracellular fluid rises upward from 180 while that of the intracellular fluid falls downward from 300 until the two meet each other After a few moments the conditions in the last line of the table obtain, with the extracellular fluid volume 18.75 liters, the intracellular volume 31 25 liters, and a milliosmolality in both compartments of 240

To calculate the changes occurring in the above example, one needs only to keep accurate accounting of the total number of milliosmoles in each fluid compartment and also in the total body water in the following manner: After the 10 liters of water are added the total body water becomes 50 liters instead of 40, but the total milliosmoles in the whole body remain the same, 12,000 Dividing 50 into 12,000, the average milliosmolar concentration in each liter of the body water under the new

Table 33–3 EFFECT OF ADMINISTERING 10 LITERS OF WATER INTRAVENOUSLY

	Extracellular			*Intracellular*			*Total Body Water*		
	Volume (Liters)	Concentration (mOsmole/L)	Total mOsmole	Volume (Liters)	Concentration (mOsmole/L)	Total mOsmole	Volume (Liters)	Concentration (mOsmole/L)	Total mOsmole
Initial	15	300	4500	25	300	7500	40	300	12000
Solution added	10	0	0	0	0	0	10	0	0
Instantaneous effect	25	180	4500	25	300	7500	50	No equilibrium	2000
After osmotic equilibrium	18 75	240	4500	31 25	240	7500	50	240	12000

conditions is found to be 240. One can readily understand that, after osmotic equilibrium has occurred throughout the body, this calculated value of 240 is the milliosmolality in both the extracellular and intracellular fluid compartments Then, dividing 240 milliosmoles into the total milliosmoles in the extracellular fluid compartment, 4500, one finds that the new volume of extracellular fluid is 18.75 liters. Dividing 240 into the total intracellular milliosmoles, 7500, gives the new intracellular fluid volume of 31 25 liters

EFFECT OF DEHYDRATION

Water can be removed from the body by evaporation from the skin, evaporation from the lungs, or excretion of a very dilute urine. In all these conditions the water leaves the extracellular fluid compartment, but on doing so some of the intracellular water passes immediately into the extracellular compartment by osmosis, thus keeping the osmolalities of the extracellular and intracellular fluids equal to each other The overall effect is called *dehydration*

EFFECT OF ADDING SALINE SOLUTION TO THE EXTRACELLULAR FLUID

If an *isotonic* saline solution is added to the extracellular fluid compartment, tha osmolality of the extracellular fluid does not change, and no osmosis results The only effect is an increase in extracellular fluid volume.

However, if a *hypertonic* solution is added to the extracellular fluid, the osmolality increases and causes osmosis of water out of the cells into the extracellular compartment.

Finally, if a *hypotonic* solution is added, the osmolality of the extracellular fluid decreases, and some of the extracellular fluid passes into the cells.

Table 33–4 gives the calculations of the effects which would occur if 2 liters of 4.4 per cent (five times isotonic concentration) sodium chloride solution were added to the extracellular fluid compartment. The added solution contains a large total number of milliosmoles, 3000, which are added to the extracellular fluid compartment. The total of milliosmoles in the body becomes 15,000, and the volume of the total body water rises from 40 to 42 Osmotic equilibration occurs within a few minutes. Dividing 42 into tha new total milliosmoles, 15,000, ona finds that the new milliosmolar concentration is 357. Now, dividing 357 into the total milliosmoles in the extracellular fluid, 7500, one finds that the new extracellular fluid volume is 21 liters And, dividing 357 into the total intracellular milliosmoles, 7500, the intracellular fluid is found also to be 21 liters.

The student should become completely familiar with this method of calculation, for an understanding of the mathematical aspects of osmotic equilibria between the two compartments is essential to the understanding of almost all fluid problems of the body

GLUCOSE AND OTHER SOLUTIONS ADMINISTERED FOR NUTRITIVE PURPOSES

Many different types of solutions are often administered intravenously to provide nutrition to patients who cannot otherwise take adequate amounts of food. Especially used are *glucose solutions* and to a lesser extent *amino acid* solutions. Rarely used are *homogenized fat solutions*. When all these are being administered, their concentrations are adjusted nearly to isotonicity, or they are given slowly enough that they do not upset the

Table 33–4 EFFECT OF ADDING 2 LITERS OF 4.4 PER CENT SODIUM CHLORIDE SOLUTION

	Extracellular			*Intracellular*			*Total Body Weight*		
	Volume (Liters)	Concentration (mOsmole/L)	Total mOsmole	Volume (Liters)	Concentration (mOsmole/L)	Total mOsmole	Volume (Liters)	Concentration (mOsmole/L)	Total mOsmole
Initial	15	300	4500	25	300	7500	40	300	12000
Solution added	2	1500	3000	0	0	0	2	1500	3000
Instantaneous effect	17	441	7500	25	300	7500	42	No equilibrium	15000
After osmotic equilibrium	21	357	7500	21	357	7500	42	357	15000

osmotic equilibria of the body fluids. However, after the glucose or other nutrient is metabolized, an excess of water often remains Ordinarily, the kidneys excrete this in the form of a very dilute urine Thus, the net result is only addition of the nutrient to the body.

Hyperalimentation. When nutritive solutions are given intravenously into a peripheral vein, their concentrations must be adjusted very nearly to isotonicity, otherwise, the abnormal osmolality will severely damage the intima of the blood vessel and the vessel will usually become plugged by a blood clot On the other hand, a patient receiving only isotonic solutions cannot tolerate enough water each day to allow total nutritive support

However, it has recently been learned that very concentrated solutions can be infused through a catheter directly into one of the major central veins where the blood flows very rapidly The solution is diluted so quickly that it will not harm the vascular system Using this procedure, enough calories can be given to a patient each 24 hours to provide full nutritive support even though only 2 to 3 liters of fluid are given to the patient each day Therefore, it is now possible to maintain patients with no intake of food by mouth whatsoever for many months at a time

REFERENCES

Adolph, E F Physiology of Man in the Desert New York Interscience Publishers 1947

Anderson D H Compartmental Modeling and Tracer Kinetics New York Springer Verlag, 1983

Andersson B Regulation of body fluids. *Annu Rev Physiol*, *39* 185 1977

Andreoli, T E, *et al* (eds) Membrane Physiology New York Plenum Press, 1980

Artz, C P, *et al* (eds) Burns, A Team Approach Philadelphia W B Saunders Co 1979

Barry, P H, and Diamond, J M Effects of unstirred layers on membrane phenomena *Physiol Rev* *64* 763 1984

Bing, D H (ed) The Chemistry and Physiology of the Human Plasma Proteins New York, Pergamon Press, 1979

Borow, M Fundamentals of Homeostasis A Clinical Approach to Fluid Electrolyte Acid Base Energy Metabolism in Health and Disease Flushing, N Y, Medical Examination Publishing Co 1977

Borut A, and Shkolnik, A Physiological adaptations to the desert environment. *Int Rev Physiol* 7 185 1974

Bradbury M W Physiology of body fluids and electrolytes *Br J Anaesthesiol*, *45* 937 1973

Burton, R F The significance of ionic concentrations in the internal media of animals. *Biol Rev*, *48* 195, 1973

Coleman, T G, *et al* Dynamics of water-isotope distribution *Am J Physiol*, *223* 1371, 1972

Dawson, K G Endocrine physiology of electrolyte metabolism *Drugs* 28(Suppl 1) 98 1984

Epstein, A N, *et al* The Neuropsychology of Thirst, Washington, V H Winston & Sons, 1973

Gilles R (ed) Mechanisms of Osmoregulation Maintenance of Cell Volume New York John Wiley & Sons, 1979

Gilmore J P Neural control of extracellular volume in the human and nonhuman primate *In* Shepherd, J T, and Abboud, F M (eds) Handbook of Physiology Sec 2, Vol III Bethesda American Physiological Society, 1983, p 885

Goldberger, E A Primer of Water, Electrolyte, and Acid Base Syndromes Philadelphia, Lea & Febiger 1980

Grinstein, S *et al* Responses of lymphocytes to anisotonic media Volume-regulating behavior *Am J Physiol*, *246* C204, 1984

Gupta, R L, *et al* Transport of Ions and Water in Animals New York Academic Press, 1977

Guyton, A C, *et al* Circulatory Physiology II Dynamics and Control of Body Fluids Philadelphia W B Saunders Co, 1975

House, C R Water Transport in Cells and Tissues. London Edward Arnold, 1974

Jamieson G A and Greenwalt, T J (eds) Blood Substitutes and Plasma Expanders New York, A R Liss, 1978

Kay R L (ed) The Physical Chemistry of Aqueous Systems New York, Plenum Publishing Corp, 1974

Klahr, S (ed) Cancer of the Prostate and Kidney New York, Plenum Publishing Corp 1983

Levick J R Blood flow and mass transport in synovial joints *In* Renkin E M and Michel, C C (eds) Handbook of Physiology Sec 2 Vol IV Bethesda American Physiological Society, 1984, p 917

Lightfoot, E N Transport Phenomena and Living Systems Biomedical Aspects of Momentum and Mass Transport New York, John Wiley & Sons 1974

Macey R I Transport of water and urea in red blood cells *Am J Physiol* *246* C195 1984

Mason E E Fluid, Electrolyte and Nutrient Therapy in Surgery Philadelphia, Lea & Febiger 1974

Parsa M H, *et al* Safe Central Venous Nutrition Springfield, Ill, Charles C Thomas, 1974

Peters T and Sjöholm I (eds) Albumin Structure, Biosynthesis Function New York Pergamon Press, 1978

Schreiber G, and Urban J The synthesis and secretion of albumin *In* Adrian, R H *et al* (eds) Reviews of Physiology, Biochemistry and Pharmacology New York Springer Verlag, 1978, p 27

Shoemaker V H, and Nagy K A Osmoregulation in amphibians and reptiles *Annu Rev Physiol* *39* 449, 1977

Smith K Fluids and Electrolytes A Conceptual Approach New York, Churchill Livingstone 1980

Solomon A K The state of water in red cells *Sci Am*, *2* 83, 1971

Vaamonde C A and Papper S Maintenance of body equilibrium *In* Frohlich, E D (ed) Pathophysiology, 2nd Ed. Philadelphia J B Lippincott Co 1976, p 265

Wilkinson A W Body Fluids in Surgery, 4th Ed New York, Churchill Livingstone, Div of Longman, 1973

Wolf A V and Crowder N A Introduction to Body Fluid Metabolism Baltimore, Williams & Wilkins, 1964

34

Formation of Urine by the Kidney: Glomerular Filtration, Tubular Function, and Plasma Clearance

The kidneys perform two major functions: First, they excrete most of the end-products of bodily metabolism and second, they control the concentrations of most of the constituents of the body fluids. The purpose of this chapter is to discuss the principles of urine formation and especially the mechanisms by which the kidneys excrete the end-products of metabolism.

PHYSIOLOGIC ANATOMY OF THE KIDNEY

The two kidneys together contain about 2,400,000 nephrons, and each nephron is capable of forming urine by itself. Therefore, in most instances, it is not necessary to discuss the entire kidney but merely the function of a single nephron to explain the function of the kidney. The nephron is composed basically of (1) a *glomerulus* from which fluid is filtered, and (2) a long *tubule* in which the filtered fluid is converted into urine on its way to the *pelvis* of the kidney.

Figure 34–1 shows the general organizational plan of the kidney, illustrating especially the distinction between the *cortex* of the kidney and the *medulla*. And Figure 34–2A illustrates the basic anatomy of the nephron, which may be described as follows: Blood enters the glomerulus through the *afferent arteriole* and then leaves through the *efferent arteriole*. The glomerulus is a network of up to 50 parallel branching and anastomosing capillaries covered by epithelial cells and encased in *Bowman's capsule*. Pressure of the blood in the glomerulus causes fluid to filter into Bowman's capsule, and from here the fluid flows into the *proximal tubule* that lies in the *cortex* of the kidney along with the glomerulus.

From the proximal tubule the fluid passes into the *loop of Henle* that dips deeply into the kidney mass, some of the loops passing all the way to the bottom of the renal medulla. Each loop is divided into the *descending limb* and the *ascending limb*. The wall of the descending limb and the lower end of the ascending limb is very thin and therefore is called the *thin segment* of the loop of Henle. However, after the ascending limb of the loop has returned part way back in the cortical direction,

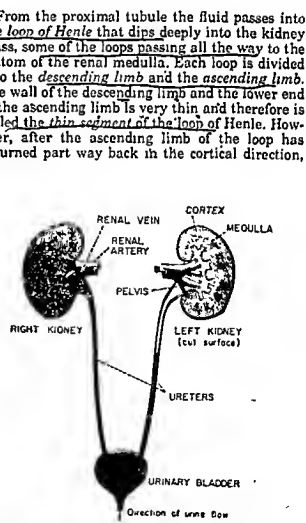
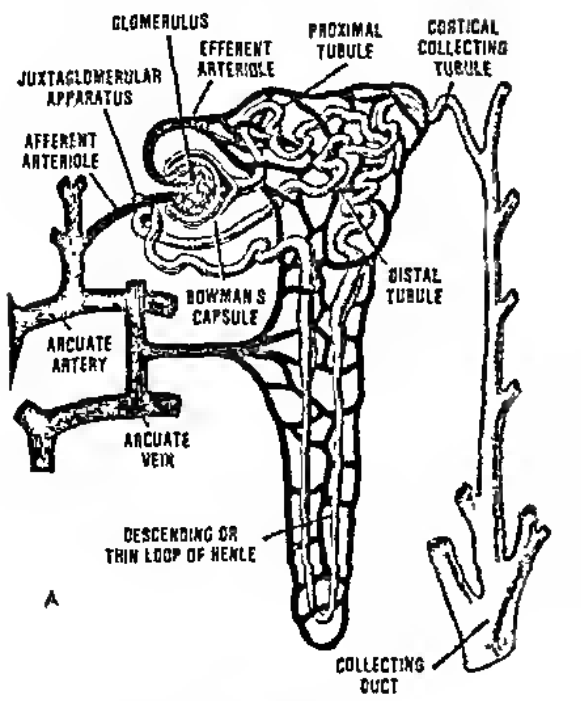


Figure 34–1. The general organizational plan of the urinary system.



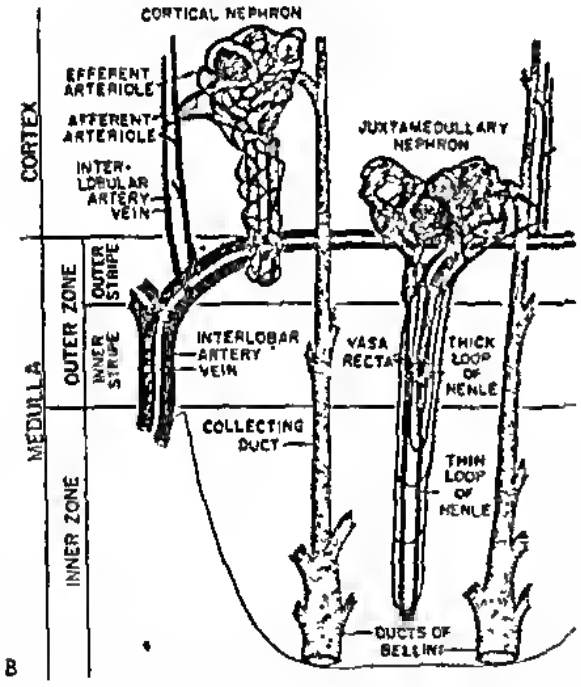


Figure 34–2. A, The nephron. (From Smith: The Kidney. Structure and Functions in Health and Disease. New York, Oxford University Press, 1951.) B, Differences between a cortical and a juxtamedullary nephron. (From Pitts: Physiology of the Kidney and Body Fluids. Chicago, Year Book Medical Publishers, 1974.)

its wall once again becomes thick like that of the other portions of the tubular system; this portion of the loop of Henle is called the *thick segment of the ascending limb*.

After passing through the loop of Henle, the [illegible] then enters the *distal tubule* which, like the [illegible] tubule, lies in the renal cortex. Then, in the cortex, as many as eight of the distal tubules coalesce to form the *collecting tubule*, the end of which turns once again away from the cortex and passes downward into the medulla, where becomes the *collecting duct*. Successive generatio of collecting ducts coalesce to form progressive larger collecting ducts that penetrate all the wa through the medulla, parallel to the loops of Henl The largest collecting ducts empty into the *ren pelvis* through the tips of the *renal papillae*; the are conical projections of the medulla that protruc into the *renal calyces*, which are themselves r cesses of the *renal pelvis*. In each kidney there a about 250 of these very large collecting ducts, eac of which transmits the urine from about 40 nephrons

As the glomerular filtrate flows through th tubules, over 99 per cent of its water and varyin amounts of its solutes are normally reabsorbe into the vascular system, and small amounts some substances are also *secreted* into the tubule The remaining tubular water and dissolved sul stances become the urine

"Cortical" and "Juxtamedullary" Nephron The characteristics of the nephrons are somewha different depending on how deep the nephrons l within the kidney mass. Those nephrons of whic the glomeruli lie close to the surface of the kidne are called *cortical nephrons*, as illustrated in Fi ure 34–2B. These have very short thin segment in their loops of Henle, and the loops penetrat only a very short distance into the outer portio of the medulla. Those nephrons of which the gl meruli lie deep in the renal cortex near the m dulla are called *juxtamedullary nephrons*. The nephrons have very long loops of Henle and esp cially long thin segments of the loops. Also, the loops of Henle penetrate deeply into the inner zon of the medulla, many of them all the way to th tips of the renal papillae.

The Peritubular Capillary Network and th Vasa Recta. Surrounding the entire tubular sy tem of the kidney is an extensive network capillaries called the *peritubular capillary networ* This network is supplied with blood from th *efferent arterioles*, blood that has already passe through the glomerulus. Most of the peritubula capillary network lies in the renal cortex alongsid the proximal tubules, distal tubules, and collectin tubules. However, from the deeper portions of thi peritubular network, long and straight capillar loops called *vasa recta* extend downward into th medulla to lie side by side with the lower parts the thin segments of the juxtaglomerular loops Henle all the way to the renal papillae. Then, lik the loops of Henle, they also loop back toward th cortex and empty into the cortical veins

Functional Diagram of the Nephron. Figur 34–3 illustrates a simplified diagram of the "phys iologic nephron." This diagram contains most o the nephron's functional structures, and it is use in the present discussion to explain many aspect of renal function.

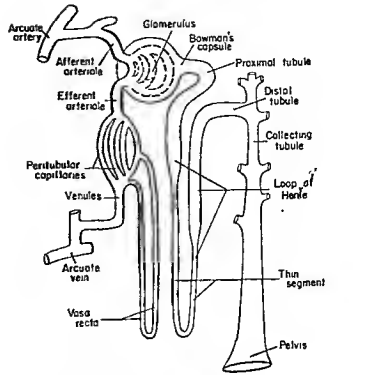


Figure 34–3. The functional nephron.

Figure 34–3. The functional nephron.

BASIC THEORY OF NEPHRON FUNCTION

The basic function of the nephron is to clean, or "clear," the blood plasma of unwanted substances as it passes through the kidney. The substances that must be cleared include particularly the end-products of metabolism, such as urea, creatinine, uric acid, and urates. In addition, many other substances, such as sodium ions, potassium ions, chloride ions, and hydrogen ions tend to accumulate in the body in excess quantities; it is the function of the nephron also to clear the plasma of these excesses.

The principal mechanism by which the nephron clears the plasma of unwanted substances is: (1) It filters a large proportion of the plasma in the flowing glomerular blood, usually about one fifth of it, through the glomerular membrane into the tubules of the nephron. (2) Then, as this filtered fluid flows through the tubules, the *unwanted substances fail to be reabsorbed* while the *wanted substances, especially the water and many of the electrolytes, are reabsorbed* back into the plasma of the peritubular capillaries. In other words, the wanted portions of the tubular fluid are returned to the blood and the unwanted portions pass into the urine.

A second mechanism by which the nephron clears the plasma of other unwanted substances is by *secretion*. That is, substances are secreted from the plasma directly through the epithelial cells lining the tubules into the tubular fluid. Thus, the urine that is eventually formed is composed mainly of *filtered* substances but also of small amounts of *secreted* substances.

RENAL BLOOD FLOW AND PRESSURES

BLOOD FLOW THROUGH THE KIDNEYS

The rate of blood flow through both kidneys of a 70 kg man is about 1200 ml/minute.

The portion of the total cardiac output that passes through the kidneys is called the *renal fraction*. Since the normal cardiac output of a 70 kg adult man is about 5600 ml/minute and the blood flow through both kidneys is about 1200 ml/minute, one can calculate that the normal renal fraction is about 21 per cent. This can vary from as little as 12 per cent to as high as 30 per cent in the normal resting person.

Special Aspects of Blood Flow Through the Renal Vasculature. Note in Figure 34–3 that there are two capillary beds associated with the nephron: (1) the *glomerulus*, and (2) the *peritubular capillaries*. The glomerular capillary bed receives its blood from the *afferent arteriole*, and from this bed the blood flows into the peritubular capillary bed through the *efferent arteriole*, which offers considerable resistance to blood flow. As a result, the glomerular capillary bed is a *high pressure bed* while the peritubular capillary bed is a *low pressure bed*. Because of the high pressure in the glomerulus, it functions in much the same way as the usual arterial ends of the tissue capillaries, with fluid filtering continually out of the glomerulus into Bowman's capsule. On the other hand, the low pressure in the peritubular capillary system causes it to function in much the same as the venous ends of the tissue c[illegible]

fluid being absorbed continually into the capillaries

Blood Flow in the Vasa Recta. A special portion of the peritubular capillary system is the vasa recta, which are a network of capillaries that descend around the lower portions of the loops of Henle These capillaries form loops in the medulla of the kidney and then return to the cortex before emptying into the veins The vasa recta play a special role in the formation of concentrated urine, which will be discussed in the following chapter

Only a small proportion of the total renal blood flow, about 1 to 2 per cent, flows through the vasa recta. In other words, blood flow through the medulla of the kidney is sluggish in contrast to the rapid blood flow in the cortex

PRESSURES IN THE RENAL CIRCULATION

Figure 34–4 gives the approximate pressures in the different parts of the renal circulation and tubules, showing an initial pressure of approximately 100 mm Hg in the large arcuate arteries and about 8 mm Hg in the veins into which the blood finally drains The two major areas of resistance to blood flow through the nephron are (1) the *small renal arteries and afferent arteriole* and (2) the *efferent arteriole* In the small arteries and afferent arteriole the pressure falls from 100 mm

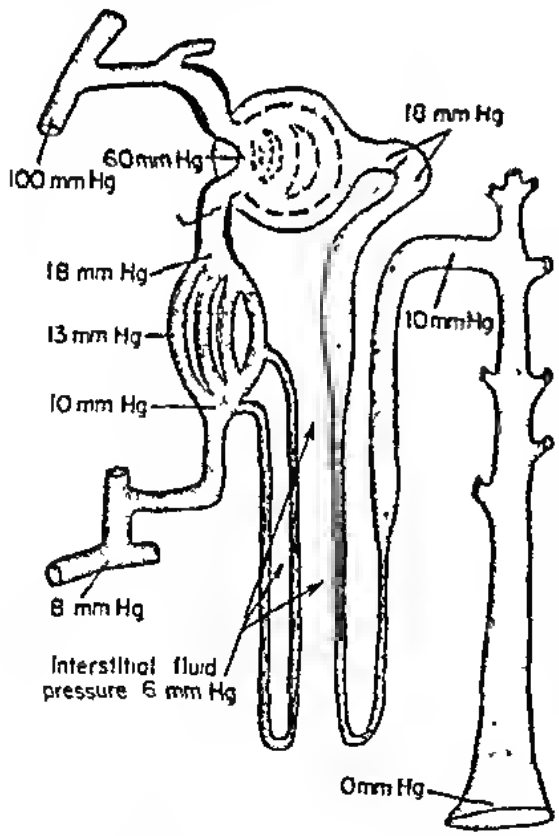


Figure 34–4. Approximate pressures at different points in the and tubules of the functional nephron and in the interstitial

Hg at its arterial end to an estimated mean pr sure of about 60 mm Hg in the glomerulus (Th pressure is still in serious doubt, having be calculated to be as high as 70 mm Hg in the d and measured to be as low as 45 mm Hg in t rat. Therefore, 60 mm Hg is only an avera estimate.) As the blood-flows through the effere arterioles from the glomerulus to the peritubul capillary system, the pressure falls another 47 m Hg to a mean peritubular capillary pressure of mm Hg .Thus, the high pressure capillary bed the glomerulus operates at a mean pressure about 60 mm Hg and therefore causes rapid filtr tion of fluid, whereas the low pressure capilla bed in the peritubular capillary system operat at a mean capillary pressure of about 13 mm H therefore allowing rapid absorption of fluid b cause of the high osmotic pressure of the plasm

"INTRARENAL PRESSURE" AND RENAL INTERSTITIAL FLUID PRESSURE

The kidney is encased in a tight *fibrous capsu* When a needle is inserted into the kidney and t pressure in the needle is gradually raised un fluid flows into the kidney tissue, the pressure which flow begins is between 10 and 18 mm H averaging perhaps 13 mm Hg. This "needle" pre sure is called the *intrarenal pressure*. It w pointed out in Chapter 30 that such needle pre sures do not measure the interstitial fluid pressu but instead measure the *total tissue pressure* (t pressure that tends to collapse blood vessels a tubules)

Recent attempts to measure the *interstitial flu pressure* of the kidney utilizing implanted perf rated capsules as described in Chapter 30 ha given an average mean value of +6 mm Hg, whic at present is probably the best estimate of inte stitial fluid pressure of the kidney.

FUNCTION OF THE PERITUBULAR CAPILLARIES

Tremendous quantities of fluid, about 180 lite each day, are filtered through all the glomeru all but 1 to 1.5 liters of this is reabsorbed from t tubules into the renal interstitial spaces a thence into the peritubular capillaries. This re resents about four times as much fluid as th reabsorbed at the venous ends of all the oth capillaries of the entire body. Therefore, one ca readily see that reabsorption of fluid into t peritubular capillaries presents a special problem However, the peritubular capillaries are extremel porous in comparison with those in other bod tissues, so extremely rapid osmosis of fluid resul ing from the colloid osmotic pressure of the plasm proteins can account for the rapid absorption th is required.

GLOMERULAR FILTRATION AND THE GLOMERULAR FILTRATE

The Glomerulor Membrane and Glomerular Permeability. The fluid that filters through the glomerulus into Bowman's capsule is called *glomerular filtrate*, and the membrane of the glomerular capillaries is called the *glomerulor membrane*. Though, in general, this membrane is similar to that of other capillaries throughout the body, it has several differences. First, it has three major layers: (1) the endothelial layer of the capillary itself, (2) a basement membrane, and (3) a layer of epithelial cells that is illustrated on the outer surfaces of the glomerular capillaries in Figure 34–5. Yet, despite the number of layers, the permeability of the glomerular membrane is from 100 to 500 times as great as that of the usual capillary

The tremendous permeability of the glomerular membrane is caused by its special structure, which is illustrated in Figure 34–5. The capillary *endothelial cells* lining the glomerulus are perforated by literally thousands of small holes called *fenestrae*. Then, outside the endothelial cells is a *basement membrane*, composed mainly of a meshwork of proteoglycan fibrillae that also has large spaces through which fluid can filter A final layer of the glomerular membrane is a layer of *epithelial cells* that line the outer surfaces of the glomerulus. However, these cells are not continuous but instead consist mainly of finger-like projections that cover the basement membrane. Figure 34–6 is a scanning electron micrograph of the outer surfaces of the glomerular capillaries, illustrating the multiple "fingers" of the epithelial cells. These fingers form slits called *slit-pores* through which the glomerular filtrate filters.

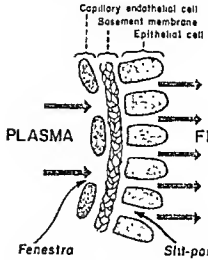


Figure 34–5. Functional structure of the glomerular membrane

Thus, the glomerular filtrate passes through three different layers before entering Bowman's capsule, but each of these layers is several hundred times as permeable as the usual capillary membrane, which accounts for the tremendous volume of glomerulor filtrate that can be formed each minute Yet, despite the tremendous permeability of the glomerulor membrane, it has an extremely high degree of selectivity for the sizes of molecules that it allows to pass.

Overall, the permeability of the glomerular membrane to substances of different molecular weights (expressed as the ratio of concentration of

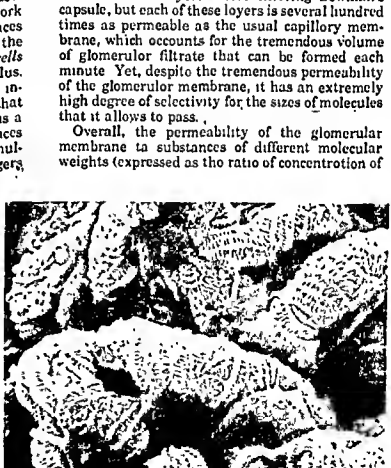


Figure 34–6. Scanning electron micrograph of a glomerulus from a normal rat kidney The visceral epithelial cells, or podocytes (P), extend multiple processes outward from the main cell body, which wrap around the individual capillary loops Note that immediately adjacent pedicels or foot processes, arise from different podocytes. (Magnification, × 3300) (from Brenner and Rector The Kidney Philadelphia, W B Saunders Company, 1981)

the dissolved substance on the filtrate side of the membrane to its concentration on the plasma side) is approximately as follows:

Molecular Weight	*Permeability*	*Example Substance*
5200	1 00	Inulin
30,000	0.5	Very small protein
69,000	0 005	Albumin

This means that at a molecular weight of 5000, the dissolved substance filters just as easily as water, but at a molecular weight of 69,000 only 0 005 per cent of the dissolved substance filters. Note that the molecular weight of the smallest plasma protein, albumin, is 69,000 Therefore, for practical purposes the glomerular membrane is almost completely impermeable to all plasma proteins but is highly permeable to essentially all other dissolved substances in normal plasma.

There are two basic reasons for the high degree of molecular selectivity by the glomerular membrane. First is the sizes of the pores in the membrane itself. That is, the pores of the membrane are large enough to allow molecules with diameters up to about 8 nanometers (80 Å) to pass through Yet, the molecular diameter of the plasma protein albumin molecule is only about 8 nanometers, which is somewhat smaller than the sizes of these large pores Therefore, why do the protein molecules not pass through in great quantity? The answer to this is the second factor that determines the permeability of the membrane: The *glomerular pores are lined with a complex of glycosylated proteins that have very strong negative electrical charges* The plasma proteins also have strong negative electrical charges. Therefore, electrostatic repulsion of the protein molecules by the pore walls keeps these molecules from passing through.

Composition of the Glomerular Filtrate. The glomerular filtrate has almost exactly the same composition as the fluid that filters from the arterial ends of the capillaries into the interstitial fluids. It contains no red blood cells and about 0.03 per cent protein, or about 1/240 the protein in the plasma.

The electrolyte and other solute composition of glomerular filtrate is also similar to that of the interstitial fluid. However, because of the paucity of negatively charged protein ions in the filtrate, a Donnan equilibrium effect occurs that causes the concentration of the other negative ions besides protein, including chloride and bicarbonate ions, to be about 5 per cent higher in the glomerular filtrate than in plasma; and the concentration of positive ions is about 5 per cent lower.

To summarize: For all practical purposes, *glomerular filtrate is the same as plasma except that no significant amount of proteins.*

THE GLOMERULAR FILTRATION RATE

The quantity of glomerular filtrate formed ea minute in all nephrons of both kidneys is call the *glomerular filtration rate.* In the normal pers this averages approximately 125 ml/min. To e press this differently, the total quantity of glome ular filtrate formed each day averages about 1 liters, or more than two times the total weight the body. Over 99 per cent of the filtrate is n mally reabsorbed in the tubules, with the remai der passing into the urine.

The Filtration Fraction. The filtration fracti is the fraction of the renal plasma flow that b comes glomerular filtrate. Since the norm plasma flow through both kidneys is 650 ml/m and the normal glomerular filtration rate in bo kidneys is 125 ml, *the average filtration fraction approximately 1/5 or 19 per cent.*

DYNAMICS OF GLOMERULAR FILTRATION

The same forces that cause fluid to filter from any high pressure capillary also apply to filtration through the glomerulus. That is, *pressure inside the glomerular capillaries* promotes filtration of fluid through the capillary membrane into Bowman's capsule. On the other hand, *colloid osmotic pressure in the blood and pressure in Bowman's capsule* oppose the filtration. (Ordinarily, the amount of protein in the filtrate in Bowman's capsule is too slight to be of any significance, but if this ever becomes increased to a significant amount, its colloid osmotic pressure will obviously also be active at the membrane, promoting increased filtration of fluid through the membrane.)

Glomerular Pressure. The glomerular pressure is the average pressure in the glomerular capillaries This unfortunately has been measured directly only in one mammal, the rat, in which the average value is about 45 mm Hg. However, from various indirect measurements it has been calculated to be from 55 to 70 mm Hg in the dog Because humans are large mammals, *a reasonable average value can be considered to be 60 mm Hg,* though, as noted later, this can increase or decrease considerably under varying conditions.

Pressure in Bowman's Capsule. In lower animals, pressure measurements actually have been made in Bowman's capsule and at different points along the renal tubules by inserting micropipets into the lumen. On the basis of these studies, *capsular pressure in the human being is estimated to be about 18 mm Hg.*

Colloid Osmotic Pressure in the Glomerular Capillaries. Because approximately one fifth of the plasma in the capillaries filters into the capsule, the protein concentration increases about 20 per cent as the blood passes from the arterial to the venous ends of the glomerular capillaries. If the normal colloid osmotic pressure of blood enter-

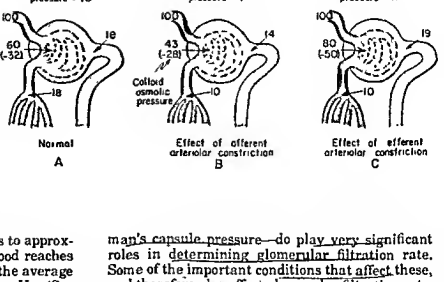


Figure 34–7. *A*, Normal pressures at different points in the nephron, and the normal filtration pressure. *B*, Effect of afferent arteriolar constriction on pressures in the nephron and on filtration pressure. *C*, Effect of efferent arteriolar constriction on pressures in the nephron and on filtration pressure.

ing the capillaries is 28 mm Hg, it rises to approximately 36 mm Hg by the time the blood reaches the venous ends of the capillaries, and the average colloid osmotic pressure is about 32 mm Hg. (See Figure 30–8 in Chapter 30, which shows the nonlinear relationship between protein concentration and colloid osmotic pressure.)

Filtration Pressure, Filtration Coefficient, and Glomerular Filtration Rate. The *filtration pressure* is the net pressure forcing fluid through the glomerular membrane, and this is equal to the *glomerular pressure* minus *the sum of glomerular colloid osmotic pressure and capsular pressure.* In Figure 34–7A, *the normal filtration pressure is shown to be about 10 mm Hg.*

The *filtration coefficient*, called K_f is defined as the glomerular filtration rate in both kidneys per mm Hg of filtration pressure. That is, the glomerular filtration rate is equal to the filtration pressure times the filtration coefficient, or

$$GFR = \text{Filtration pressure} \cdot K_f$$

The normal filtration coefficient is 12.5 ml/min/mm Hg of filtration pressure. Thus, at a normal mean filtration pressure of 10 mm Hg, the total filtration rate of both kidneys is 125 ml/min.

(Note: From studies in rats, several investigators have postulated that the filtration coefficient is much greater than given here and that the filtration pressure is much lower, but these values are still under discussion.)

FACTORS THAT AFFECT THE GLOMERULAR FILTRATION RATE

It is clear from the above equation that the filtration pressure and the filtration coefficient determine the glomerular filtration rate. The filtration coefficient probably does not change greatly from normal except when the kidneys become diseased.

On the other hand, the three factors that determine filtration pressure—(1) glomerular pressure, (2) plasma colloid osmotic pressure, and (3) Bowman's capsule pressure—do play very significant roles in determining glomerular filtration rate. Some of the important conditions that affect these, and therefore also affect glomerular filtration rate, are the following:

Effect of Renal Blood Flow on Glomerular Filtration Rate. An increase in the rate of blood flow through the nephrons greatly increases the glomerular filtration rate. One of the reasons for this is that the increasing flow increases the glomerular pressure, which obviously enhances filtration. However, a second reason that is not quite so easily understood is the following: At normal renal blood flow, about 20 per cent of the plasma is filtered through the glomerular membrane; therefore, the concentration of the plasma proteins becomes considerably increased before the blood leaves the glomerulus and, consequently, the colloid osmotic pressure increases. This rising colloid osmotic pressure exerts a strong influence on reducing further filtration.

Now let us see what effect increased blood flow will have on this. As the flow increases, greater quantities of plasma enter the glomeruli so that filtration of fluid from the plasma causes a smaller percentage increase in protein concentration and colloid osmotic pressure. Therefore, the rising colloid osmotic pressure now exerts far less inhibitory influence on glomerular filtration. Consequently, even when the glomerular pressure remains constant, the greater the rate of blood flow into the glomerulus, the greater the glomerular filtration rate.

Effect of Afferent Arteriolar Constriction on Glomerular Filtration Rate. Afferent arteriolar constriction decreases the rate of blood flow into the glomerulus and also decreases the glomerular pressure, both of these effects decreasing the filtration rate. This effect is illustrated in Figure 34–7B. Conversely, dilatation of the afferent arteriole increases the glomerular pressure, with a corresponding increase in glomerular filtration rate.

Effect of Efferent Arteriolar Constriction. Constriction of the efferent arteriole increases the resistance to outflow from the glomeruli. This

the dissolved substance on the filtrate side of the membrane to its concentration on the plasma side) is approximately as follows:

Molecular Weight	Permeability	Example Substance
5200	1.00	Inulin
30,000	0.5	Very small protein
69,000	0.005	Albumin

This means that at a molecular weight of 5000, the dissolved substance filters just as easily as water, but at a molecular weight of 69,000 only 0.005 per cent of the dissolved substance filters. Note that the molecular weight of the smallest plasma protein, albumin, is 69,000. Therefore, for practical purposes the glomerular membrane is almost completely impermeable to all plasma proteins but is highly permeable to essentially all other dissolved substances in normal plasma.

There are two basic reasons for the high degree of molecular selectivity by the glomerular membrane. First is the sizes of the pores in the membrane itself. That is, the pores of the membrane are large enough to allow molecules with diameters up to about 8 nanometers (80 Å) to pass through. Yet, the molecular diameter of the plasma protein albumin molecule is only about 6 nanometers, which is somewhat smaller than the sizes of these large pores. Therefore, why do the protein molecules not pass through in great quantity? The answer to this is the second factor that determines the permeability of the membrane. The *glomerular pores are lined with a complex of glycosylated proteins that have very strong negative electrical charges.* The plasma proteins also have strong negative electrical charges. Therefore, electrostatic repulsion of the protein molecules by the pore walls keeps these molecules from passing through.

Composition of the Glomerular Filtrate. The glomerular filtrate has almost exactly the same composition as the fluid that filters from the arterial ends of the capillaries into the interstitial fluids. It contains no red blood cells and about 0.03 per cent protein, or about 1/240 the protein in the plasma.

The electrolyte and other solute composition of glomerular filtrate is also similar to that of the interstitial fluid. However, because of the paucity of negatively charged protein ions in the filtrate, a Donnan equilibrium effect occurs that causes the concentration of the other negative ions besides protein, including chloride and bicarbonate ions, to be about 5 per cent higher in the glomerular filtrate than in plasma; and the concentration of positive ions is about 5 per cent lower.

To summarize: For all practical purposes, *glomerular filtrate is the same as plasma except that no significant amount of proteins.*

THE GLOMERULAR FILTRATION RATE

The quantity of glomerular filtrate formed each minute in all nephrons of both kidneys is called the *glomerular filtration rate.* In the normal person this averages approximately 125 ml/min. To express this differently, the total quantity of glomerular filtrate formed each day averages about 180 liters, or more than two times the total weight of the body. Over 99 per cent of the filtrate is normally reabsorbed in the tubules, with the remainder passing into the urine.

The Filtration Fraction. The filtration fraction is the fraction of the renal plasma flow that becomes glomerular filtrate. Since the normal plasma flow through both kidneys is 650 ml/min and the normal glomerular filtration rate in both kidneys is 125 ml, *the average filtration fraction is approximately 1/5 or 19 per cent.*

DYNAMICS OF GLOMERULAR FILTRATION

The same forces that cause fluid to filter from any high pressure capillary also apply to filtration through the glomerulus. That is, *pressure inside the glomerular capillaries* promotes filtration of fluid through the capillary membrane into Bowman's capsule. On the other hand, *colloid osmotic pressure in the blood and pressure in Bowman's capsule* oppose the filtration. (Ordinarily, the amount of protein in the filtrate in Bowman's capsule is too slight to be of any significance, but if this ever becomes increased to a significant amount, its colloid osmotic pressure will obviously also be active at the membrane, promoting increased filtration of fluid through the membrane.)

Glomerular Pressure. The glomerular pressure is the average pressure in the glomerular capillaries. This unfortunately has been measured directly only in one mammal, the rat, in which the average value is about 45 mm Hg. However, from various indirect measurements it has been calculated to be from 55 to 70 mm Hg in the dog. Because humans are large mammals, *a reasonable average value can be considered to be 60 mm Hg,* though, as noted later, this can increase or decrease considerably under varying conditions.

Pressure in Bowman's Capsule. In lower animals, pressure measurements actually have been made in Bowman's capsule and at different points along the renal tubules by inserting micropipets into the lumen. On the basis of these studies, *capsular pressure in the human being is estimated to be about 18 mm Hg.*

Colloid Osmotic Pressure in the Glomerular Capillaries. Because approximately one fifth of the plasma in the capillaries filters into the capsule, the protein concentration increases about 20 per cent as the blood passes from the arterial to the venous ends of the glomerular capillaries. If the normal colloid osmotic pressure of blood enter-

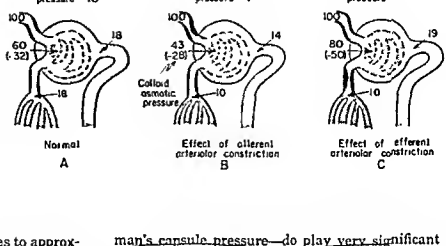


Figure 34–7. A, Normal pressures at different points in the nephron, and the normal filtration pressure. B, Effect of afferent arteriolar constriction on pressures in the nephron and on filtration pressure. C, Effect of efferent arteriolar constriction on pressures in the nephron and on filtration pressure

ing the capillaries is 28 mm Hg, it rises to approximately 36 mm Hg by the time the blood reaches the venous ends of the capillaries, and the average colloid osmotic pressure is about 32 mm Hg. (See Figure 30–8 in Chapter 30, which shows the nonlinear relationship between protein concentration and colloid osmotic pressure.)

Filtration Pressure, Filtration Coefficient, and Glomerular Filtration Rate. The *filtration pressure* is the net pressure forcing fluid through the glomerular membrane, and this is equal to the *glomerular pressure* minus *the sum of glomerular colloid osmotic pressure and capsular pressure.* In Figure 34–7A, *the normal filtration pressure is shown to be about 10 mm Hg.*

The *filtration coefficient,* called K_f is defined as the glomerular filtration rate in both kidneys per mm Hg of filtration pressure. That is, the glomerular filtration rate is equal to the filtration pressure times the filtration coefficient, or

$$GFR = \text{Filtration pressure} \cdot K_f$$

The normal filtration coefficient is 12.5 ml/min/mm Hg of filtration pressure. Thus, at a normal mean filtration pressure of 10 mm Hg, the total filtration rate of both kidneys is 125 ml/min.

(Note: From studies in rats, several investigators have postulated that the filtration coefficient is much greater than given here and that the filtration pressure is much lower, but these values are still under discussion.)

FACTORS THAT AFFECT THE GLOMERULAR FILTRATION RATE

It is clear from the above equation that the filtration pressure and the filtration coefficient determine the glomerular filtration rate. The filtration coefficient probably does not change greatly from normal except when the kidneys become diseased.

On the other hand, the three factors that determine filtration pressure—(1) glomerular pressure, (2) plasma colloid osmotic pressure, and (3) Bowman's capsule pressure—do play very significant roles in determining glomerular filtration rate. Some of the important conditions that affect these, and therefore also affect glomerular filtration rate, are the following:

Effect of Renal Blood Flow on Glomerular Filtration Rate. An increase in the rate of blood flow through the nephrons greatly increases the glomerular filtration rate. One of the reasons for this is that the increasing flow increases the glomerular pressure, which obviously enhances filtration. However, a second reason that is not quite so easily understood is the following: At normal renal blood flow, about 20 per cent of the plasma is filtered through the glomerular membrane; therefore, the concentration of the plasma proteins becomes considerably increased before the blood leaves the glomerulus and, consequently, the colloid osmotic pressure increases. This rising colloid osmotic pressure exerts a strong influence on reducing further filtration.

Now let us see what effect increased blood flow will have on this. As the flow increases, greater quantities of plasma enter the glomerulus so that filtration of fluid from the plasma causes a smaller percentage increase in protein concentration and colloid osmotic pressure. Therefore, the rising colloid osmotic pressure now exerts far less inhibitory influence on glomerular filtration. Consequently, even when the glomerular pressure remains constant, the greater the rate of blood flow into the glomerulus, the greater the glomerular filtration rate.

Effect of Afferent Arteriolar Constriction on Glomerular Filtration Rate. Afferent arteriolar constriction decreases the rate of blood flow into the glomerulus and also decreases the glomerular pressure, both of these effects decreasing the filtration rate. This effect is illustrated in Figure 34–7B. Conversely, dilatation of the afferent arteriole increases the glomerular pressure, with a corresponding increase in glomerular filtration rate.

Effect of Efferent Arteriolar Constriction. Constriction of the efferent arteriole increases the resistance to outflow from the glomeruli. This

obviously increases the glomerular pressure and at small increases in efferent resistance often causes o slight increase in glomerular filtration rate, as illustrated in Figure 34–7C. *However*, the blood flow decreases at the same time, and if the degree of efferent arteriolar constriction is moderate or severe, the plasma will remain for a long period of time in the glomerulus, and extra large portions of plasma will filter out. This will increase the plasma colloid osmotic pressure to excessive levels, which will cause a paradoxical decrease in the glomerular filtration rate despite the elevated glomerulor pressure.

Effect of Sympathetic Stimulotion. During sympathetic stimulation of the kidneys, the afferent arterioles are constricted preferentiolly, thereby decreasing the glomerular filtration rate. With very strong sympothetic stimulation, glomerular blood flow and the glomerular pressure are reduced so greatly that glomerular filtration can decrease to only a few per cent of normal, and the urinary output actually can fall to zero for as long as 5 to 10 minutes.

Effect of Arterial Pressure. One would expect an increase in arterial pressure to cause a proportionate increase in all pressures in the nephron and therefore to increase the glomerular filtration rate to a great extent. In actual fact, this effect is greatly blunted because of a phenomenon called *autoregulation*, which is explained in the following chapter Briefly, when the arterial pressure rises, afferent arteriolar constriction occurs automatically; this prevents a significant rise in glomerular pressure despite the rise in arterial pressure. Therefore, the glomerular filtration rate usually increases only a few per cent even when the mean arterial pressure rises from its normal value of 100 mm Hg to as high as 150 mm Hg

Nevertheless, we shall also see in the following chapter that even a small percentage increase in glomerular filtration rate can cause a manyfold increase in urinary output Therefore, on increase in orterial pressure can greatly increase urinary output even though it affects glomerular filtration rate only slightly.

REABSORPTION AND SECRETION IN THE TUBULES

The glomerular filtrate entering the tubules of the nephron flows (1) through the *proximal tubule*, (2) through the *loop of Henle*, (3) through the *distal tubule*, (4) through the *collecting tubule*, and (5) through the *collecting duct* into the pelvis of the kidney. Along this course, substances ore selectively reabsorbed or secreted by the tubular epithelium, and the resultant fluid entering the pelvis urine. Reabsorption plays a much greater role ... does secretion in this formation of urine, but ...tion is especially important in determining the amounts of potassium ions, hydrogen ions, and a few other substances in the urine, as discussed later.

Ordinarily, more than 99 per cent of the water in the glomerular filtrate is reabsorbed as it passes through the tubules. Therefore, if some dissolved constituent of the glomerular filtrate is not reabsorbed at all along the entire course of the tubules, this reabsorption of water obviously concentrates the substance more than 99-fold. On the other hand, some constituents, such as glucose and amino acids, are reabsorbed almost entirely so that their concentrations decrease almost to zero before the fluid becomes urine. In this way the tubules separate substances that are to be conserved in the body from those that are to be eliminated in the urine.

BASIC MECHANISMS OF ABSORPTION AND SECRETION IN THE TUBULES

The basic mechanisms for transport through t tubular membrane are essentially the same those for transport through other membranes the body. These can be divided into *octive transp* and *passive transport*. The basic essentials of the mechanisms are described here, but addition details can be found in Chapter 9.

Active Transport Through the Tubular Membrane

Transport of Sodium Ions. Figure 34–8 illustrates, by way of example, the mechanism f active transport of sodium from the lumen of t proximal tubule into the peritubular capillar Note, first, the character of the epithelial cells th line the tubule Each epithelial cell has a "brus border on its luminal surface. This brush is co posed of literally thousands of very minute mic

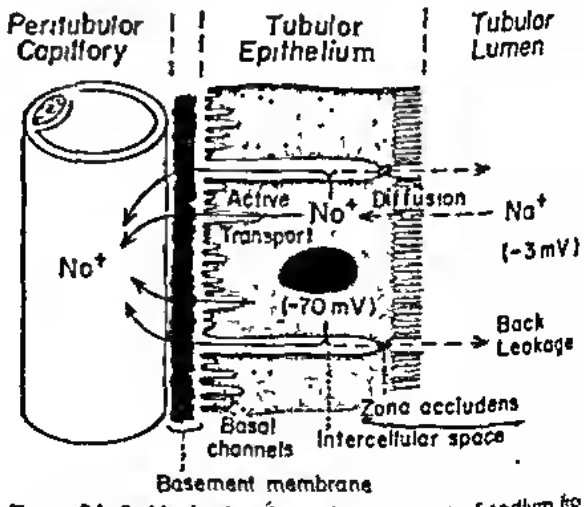


Figure 34–8. Mechanism for active transport of sodium fro the tubular lumen into the peritubular capillary, illustrating acti transport at the base and sides of the epithelial cell and diffus through the luminal border of the cell

villi that multiply the surface area of luminal exposure of the cell about twentyfold. The base of the cell sits on the basement membrane, but it is pocked by an extensive system of *basal channels* that multiply the basal and lateral surface areas manyfold. The epithelial cells are attached to each other only near the brush border, a tight attachment area called either a *tight junction* or a *zonula occludens*.

Active transport of sodium occurs from inside the epithelial cell through its basolateral membranes into the interstitial spaces of the kidney. This transport outward from the cell diminishes the sodium concentration inside the cell. Then, because this low concentration inside the cell establishes a sodium ion concentration gradient from the lumen of the tubule to the inside of the cell, sodium ions diffuse from the tubule through the brush border into the cell, and this sodium, too, is carried the rest of the way into the interstitial spaces by the sodium pump.

Note in Figure 34–8 that the electrical potential inside the epithelial cell, caused by the continual active transport of the sodium out of the cells, is approximately—70 millivolts. This very negative intracellular voltage is important because it, as well as the low concentration of sodium inside the epithelial cell, is an additional important factor that causes sodium diffusion from the tubular lumen into the cell. These two factors together are called the "electrochemical" gradient. The rapid diffusion of sodium into the cell is also facilitated by a very high permeability of the brush border membrane to sodium, primarily because of the extensive surface area of the thousands of microvilli. However, another factor that increases this permeability is that sodium ions bind with a *sodium carrier protein* in the membrane that enhances the movement of the sodium through the membrane by the process called *facilitated diffusion*, which was explained in Chapter 9.

Transport of Other Substances from the Tubules. Other substances besides sodium that are actively absorbed through the tubular epithelial cells include *glucose, amino acids, calcium ions, potassium ions, chloride ions, bicarbonate ions, phosphate ions, urate ions*, and others. As was explained in Chapter 9, both glucose and the amino acids are transported from the tubular lumen through the brush border by a process called *sodium cotransport*. That is, the glucose or amino acid molecule binds with the same sodium carrier molecule in the brush border that transports sodium ions through this membrane. Then, as the sodium diffuses inward through the membrane, it literally pulls the glucose or amino acid along with it. Once inside the epithelial cell, the sodium and the glucose or amino acid split from the carrier. The glucose or amino acid then diffuses through the basal membrane of the cell and thence into the peritubular capillaries by the process of facilitated diffusion, as was discussed in Chapter 9.

Active Secretion into the Tubules. In addition, some substances are actively *secreted* into all or some portions of the tubules; these include especially *hydrogen ions, potassium ions*, and *urate ions*. Active secretion occurs in the same way as active absorption except that the cell membrane transports the secreted substance in the opposite direction; for some of the actively secreted substances, it is the brush border rather than the basolateral membrane of the cell that provides the active transport mechanism.

Passive Absorption of Water: Osmosis Through the Tubular Epithelium

When the different solutes are transported out of the tubule, their total concentration decreases inside the tubular lumen but increases on the outside in the renal interstitium. This obviously creates a concentration difference that causes osmosis of water in the same direction that the solutes have been transported.

However, some portions of the tubular system are far more permeable to water than are others. In those portions that are highly permeable, such as the proximal tubules, osmosis of water occurs so rapidly that the osmolar concentration of solutes on the interstitial side of the membrane is almost never more than a few milliosmoles greater than on the intratubular side. For instance, referring again to Figure 34-8, when sodium ions are transported from a tubule into the peritubular fluid of the intercellular spaces and basal channels, the osmotic pressure in this fluid increases while that in the tubule decreases. Since the proximal tubular membrane is highly permeable to water, osmosis occurs almost instantaneously into the peritubular fluid to re-equilibrate the osmolar concentrations on the two sides of the membrane.

On the other hand, portions of the distal tubule are an example of a tubular area that is almost completely impermeable to water, a fact that plays a very important role in the mechanism for controlling urine concentration. This will be discussed later.

Passive Absorption of Urea and Other Nonactively Transported Solutes by the Process of Diffusion

When water is reabsorbed by osmosis, about half the urea in the tubular fluid remains behind; therefore, the concentration of urea in the tubular fluid rises, which obviously establishes a concentration difference for urea between the tubular and interstitial fluids. This in turn causes urea also to diffuse from the tubular fluid into the interstitium. And this same effect also occurs for other tubular solutes that are not actively reabsorbed but that are diffusible through the tubular membrane.

The rate of resorption of a nonactively reab-

sorbed solute is determined by (1) the amount of water that is reabsorbed, because this determines the tubular concentration of the solute, and (2) the permeability of the tubular membrane for the solute. The permeability of the membrane for urea in most parts of the tubules is far less than that for water, which means that far less urea is reabsorbed than water. Therefore, a large proportion of the urea remains in the tubules and is lost in the urine—usually over 50 per cent of all that enters the glomerular filtrate. The permeability of the tubular membrane for reabsorption of creatinine, inulin (a large polysaccharide), mannitol (a monosaccharide), and sucrose is zero, which means that once these substances have filtered into the glomerular filtrate, 100 per cent of that which enters the glomerular filtrate passes on into the urine.

Diffusion Caused by Electrical Differences Across the Tubular Membrane: The active absorption of sodium from the tubule creates negativity inside the tubular lumen with respect to the interstitial fluid, as was illustrated in Figure 34-8 In the early proximal tubule, this electrical potential is approximately −3 millivolts. In the distal end of the distal tubule, it ranges from −10 to −70 millivolts. In some other segments of the tubules, this potential is a few millivolts positive—for instance, in the so-called "diluting segment" of tha dietal tubule, as will be discussed in more detail later in this chapter.

When the electrical potential is negative inside the tubule, this repels negative ions, such as chloride and phosphate ions, causing them to diffuse out of tha tubules and into the interstitial fluid (if the tubular membrane is permeable to the ions). On the other hand, it will attract positive ions, such as sodium and potssium ions, from the interstitial fluid into the tubules. Conversely, when the intratubular fluid is positive, positive ions then move toward the interstitial fluid and negative ions toward the tubular lumen.

ABSORPTIVE CAPABILITIES OF DIFFERENT TUBULE SEGMENTS

In subsequent sections of this chapter the absorption and secretion of specific substances in different segments of the tubular system will be discussed. However, it is important first to point out basic differences between the absorptive a secretory capabilities of the different types of bules.

Proximal Tubular Epithelium. Figure 34 illustrates the celluiar characteristics of the bular membrane in (1) the proximal tubule, the thin segment of the loop of Henle, (3) the dis tubule, and (4) the collecting duct. The proxin tubular cells have the appearance of being higi metabolic cells, having iarge numbers of mitoch dria to support extremely rapid active transp processes; and, true enough, one finds that ab *65 per cent of all reabsorption and secretion t occur in the tubular system take place in the pr imal tubules.* That is, 65 per cent of the glomeru filtrate normally is reabsorbed before reaching t loops of Henie. As already described in relation sodium transport, the proximal tubular epithel cells have an extensive brush border, They a have a far-ranging labyrinth of intercellular a basal channels that provide an extensive me brane area on the interstitial fluid side of t epithelium, the side where abundant active tra port of sodium ions occurs

Another special characteristic of the proxin tubule epithelium that is different from the e thelium elsewhere in the tubular system is t following: The so-called "tight" junctions of t proximal tubules that attach the epithelial ce together are not very tight. Both water molecul as well as many small molecular weight solu can leak through these junctions. Therefore, if e too much fluid has been absorbed from the pro mal tubular lumen by the epithelial cells and t pressure of this fluid has become excessively e vated in the interstitial spaces on the outside the tubule, then large portions of the fluid w leak back again into the tubular lumen. For t reason, the net rate of absorption from the pro mal tubules is determined not only by the rate active transport through the tubular epithelio but also by all factors that affect the intrare interstitial fluid pressure, a high interstitial pr sure decreasing net absorption while a low pr sure enhances absorption. For instance, when t pressure in the peritubular capillaries is low, when there is a high plasma colloid osmotic pr sure, either of these will promote rapid absorpti of fluid from the interstitium and therefore preve back leakage of fluid into the proximal tubul

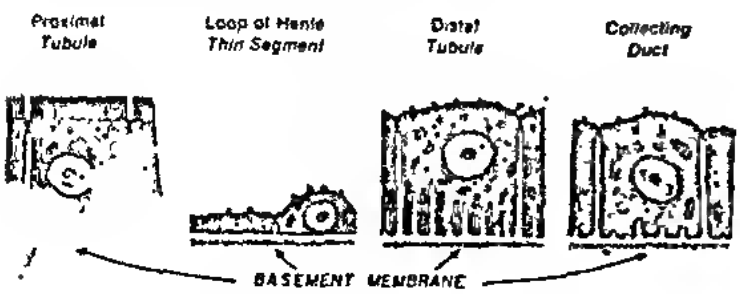


Figure 34–9. Characteristics of the epithelial cells different tubular segments.

lumen. In effect, therefore, either of these two—low capillary pressure or high plasma colloid osmotic pressure—will increase the net rate of proximal tubular reabsorption.

Thin Segment of the Loop of Henle. The epithelium of the thin segment of the loop of Henle, as the name implies, is very thin. The cells have no brush border and very few mitochondria, indicating a minimal level of metabolic activity. The *descending portion* of this thin segment is highly permeable to water and moderately permeable to urea, sodium, and most other ions. Therefore, it appears to be adapted primarily for simple diffusion of substances through its walls.

The ascending portion of the thin segment, on the other hand, is believed to be different in one very important characteristic; it is believed to be far less permeable to water but more permeable to urea than is the descending portion. These differences are important for explaining the function of the countercurrent mechanism for concentrating urine that will be discussed in the following chapter.

Thick Segment of the Loop of Henle. The thick segment of the loop of Henle begins in the ascending limb of the loop where the epithelial cells become grossly thickened, as illustrated in Figure 34–2. This segment then ascends all the way back to the same glomerulus from which the tubule originated and passes snugly through the angle between the afferent and efferent arterioles, forming a complex with these arterioles called the *juxtaglomerular complex*. This structure will be discussed in greater detail in the following chapter because it plays very important roles in controlling nephron function.

The epithelial cells of the thick segment of the loop of Henle are similar to those of the proximal tubules except that they have a rudimentary brush border, fewer basal channels, and much tighter "tight" junctions where the cells attach to each other. The cells are especially adapted for strong active transport of sodium and potassium ions and perhaps chloride ions as well, transporting these from the tubular lumen into the interstitial fluid.

On the other hand, this thick segment is almost entirely impermeable to both water and urea. Therefore, even though about three quarters of all the ions in the tubular fluid are transported out of the thick segment into the interstitium, almost all the water and urea remain in the tubule. Thus, the ascending limb tubular fluid becomes very dilute, except that its concentration of urea is high. We shall see in the discussions of the following chapter that this thick segment plays an exceedingly important role in controlling the degree of dilution or concentration of the urine that is eventually formed by the kidney.

The Distal Tubule. The distal tubule begins at the juxtaglomerular complex, continuing from the thick segment of the ascending limb of the loop of Henle. The distal tubule is highly convoluted and eventually coalesces with several other distal tubules to form the cortical collecting tubule that was described earlier.

The distal tubule is divided into two important functional segments, the *diluting segment* and the *late distal tubule*.

The Diluting Segment. The first half of the distal tubule has almost the same characteristics as the thick segment of the ascending limb of the loop of Henle. It absorbs most of the ions avidly but is almost entirely impermeable to both water and urea. Therefore, this diluting segment, too, contributes to the dilution of the tubular fluid in the same way as the thick segment of the ascending limb of the loop of Henle.

The Late Distal Tubule and Cortical Collecting Tubule. The functional characteristics of the late distal tubule and the cortical collecting tubule are similar to each other; even their lining epithelial cells are similar. Several important characteristics of these tubular segments are the following:

(1) The epithelium of both is almost entirely impermeable to urea, the same as in the diluting segment of the distal tubule, so that essentially all the urea passes on into the collecting duct eventually to be excreted in the urine.

(2) These two segments reabsorb *sodium ions* avidly, but the rate of this reabsorption is controlled to a very great extent by aldosterone, as we shall discuss in more detail in the following chapter. Simultaneously with the pumping of sodium out of the tubular lumen into the peritubular interstitium, *potassium ions* are transported in the opposite direction into the tubular lumen; this too is controlled by aldosterone and by several other factors including the concentration of potassium ions in the body fluids. Thus, potassium ions are actively secreted into these tubular segments, and *it is mainly by this means that the potassium ion concentration is controlled in the extracellular fluids of the body.*

(3) The late distal tubule and the cortical collecting tubule differ from the diluting segment in still another very important aspect: they are *permeable to water in the presence of antidiuretic hormone* but *impermeable when this hormone is absent*, providing a method for controlling the degree of dilution of the urine, a subject that we shall discuss in detail in the following chapter. The collecting duct also shares this responsiveness to antidiuretic hormone.

The Collecting Duct Epithelium. The epithelial cells of the collecting duct are nearly cuboidal in shape, with smooth surfaces, and contain relatively few mitochondria. This epithelium has two especially important characteristics for renal function:

(1) The permeability of the collecting duct to water is controlled mainly by the level of antidiuretic hormone in the circulating blood, as just mentioned. When excessive amounts of antidiuretic hormone are available, water is reabsorbed

into the medullary interstitium in great quantities, thus reducing the volume of urine and concentrating most of the dissolved substances in the urine. The collecting duct epithelium is slightly permeable to urea as well, and this permeability increases slightly in the presence of antidiuretic hormone. Therefore, some urea is reabsorbed into the medullary interstitium. Most of this then diffuses back into the loop of Henle and is returned once again to the collecting duct and is finally secreted.

(2) The second important characteristic of the collecting duct epithelium is its capability for secreting hydrogen ions against a very high hydrogen ion gradient. Therefore, as will be explained in Chapter 37, the collecting duct plays an exceedingly important role in controlling the acid-base balance of the body fluids.

REABSORPTION AND SECRETION OF INDIVIDUAL SUBSTANCES IN DIFFERENT SEGMENTS OF THE TUBULES

Reabsorption of Water; Rate of Fluid Flow at Different Points in the Tubular System. Water transport occurs entirely by osmotic diffusion. That is, whenever some solute in the glomerular filtrate is absorbed either by active reabsorption or by diffusion caused by an electrochemical gradient, the resulting decreased concentration of solute in the tubular fluid and increased concentration in the interstitial fluid causes osmosis of water out of the tubules. Therefore, the volume of tubular fluid progressively diminishes along the tubular system.

Figure 34–10 depicts the rates of fluid flow axially along the tubular system at different points in the tubules. In the two kidneys of the human being, the total fluid volumes flowing into each segment each minute (under normal resting conditions) are the following:

	ml/min
Glomerular filtrate	125
Flowing into the loops of Henle	45
Flowing into the distal tubules	25
Flowing into the collecting tubules	12
Flowing into the urine	1

From this chart one can also deduce the per cent of the glomerular filtrate water that is reabsorbed in each segment of the tubules, as follows:

	Per Cent
Proximal tubules	65
Loops of Henle	15
Distal tubules	10
Collecting tubules and ducts	9.3
Passing into the urine	0.7

We shall see later in this and the following chapters that these values vary greatly under

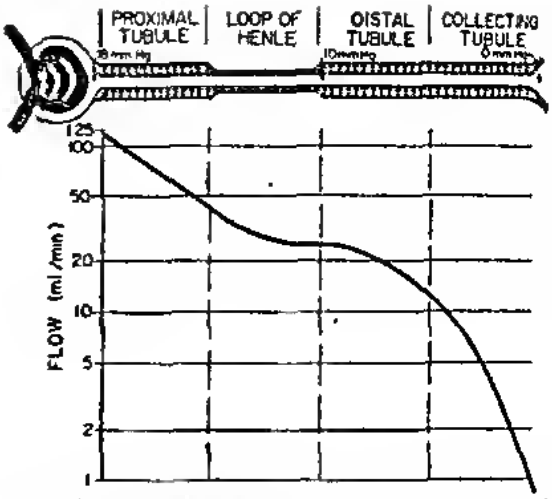


Figure 34–10. Volume flow of fluid in each segment of the tubular system per minute. Note that the flow is plotted on a semi logarithmic scale illustrating the tremendous difference in flow between the earlier and later segments of the tubules.

different operational conditions of the kidney, particularly when the kidney is forming very dilute or very concentrated urine.

Reabsorption of Substances of Nutritional Value to the Body—Glucose, Proteins, Amino Acids, Acetoacetate Ions, and Vitamins. Five different substances in the glomerular filtrate of particular importance to bodily nutrition are *glucose, proteins, amino acids, acetoacetate ions,* and *the vitamins.* Normally all of these are completely or almost completely *reabsorbed by active processes in the proximal tubules of the kidney,* as shown in Figure 34–11 for glucose, protein, and amino acids.

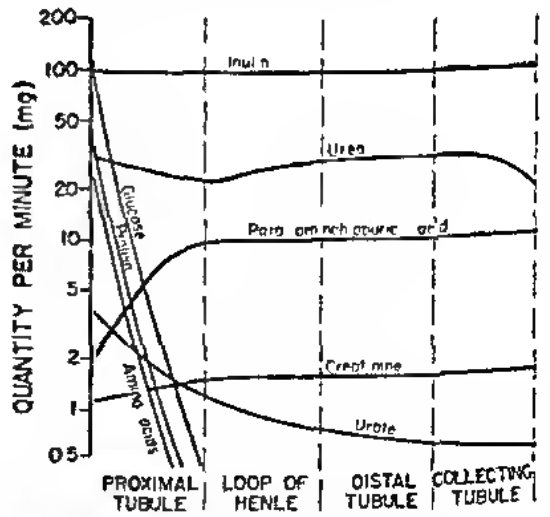


Figure 34–11. Rates of flow of important organic substances through the tubules at each point along the tubular course. (Two of these substances, inulin and para aminohippuric acid, are not normally present but are important test substances.) This figure demonstrates reabsorption of all nutritionally important substances in the proximal tubules and poor reabsorption of the metabolic end-products in all segments of the tubules. Note the total absence of reabsorption of inulin and the secretion of para aminohippuric acid into the proximal tubules.

Therefore, almost none of these substances remain in the tubular fluid entering the loop of Henle.

Special Mechanism for Absorption of Protein. As much as 30 grams of plasma protein filter into the glomerular filtrate each day. This would be a great metabolic drain on the body if the protein were not returned to the body fluids. Because the protein molecule is much too large to be transported by the usual transport processes, *protein is absorbed through the brush border of the proximal tubular epithelium by pinocytosis*, which means simply that the protein attaches itself to the membrane and this portion of the membrane then invaginates to the interior of the cell. Once inside the cell the protein is digested into its constituent amino acids, which are then absorbed by facilitated diffusion through the base and sides of the cell into the interstitial fluid. Details of the pinocytosis mechanism were discussed in Chapter 9.

Poor Reabsorption of the Metabolic End-Products: Urea, Creatinine, and Others. Figure 34–11 also illustrates the rates of flow of three major metabolic end-products in the different segments of the tubular system—urea, creatinine, and urate ions. Note, especially, that only moderate quantities of *urea*—about 50 per cent of the total—are reabsorbed during the entire course through the tubular system.

Creatinine is not reabsorbed in the tubules at all; indeed, small quantities of creatinine are actually secreted into the tubules by the proximal tubules so that the total quantity of creatinine increases about 20 per cent.

The *urate ion* is absorbed much more than urea—about 86 per cent reabsorption. But even so, large quantities of urate still remain in the fluid that finally issues into the urine. Several other end-products, such as *sulfates, phosphates,* and *nitrates,* are transported in much the same way as urate ions. All these are normally reabsorbed to a far less extent than is water so that their concentrations become greatly increased as they flow along the tubules. Yet, *each is actively reabsorbed to some extent,* which keeps their concentrations in the extracellular fluid from ever falling too low.

Transport of Inulin and Para-Aminohippuric Acid by the Tubules. Note also in Figure 34–11 that when the substance *inulin,* a large polysaccharide, is infused into the blood and is then filtered in the glomerular filtrate, its flow rate remains exactly the same throughout the entire tubular system. The cause of this is simply that inulin is neither reabsorbed nor secreted in any segment of the tubules.

Finally, Figure 34–11 shows that when *para-aminohippuric acid* (PAH) is infused into the blood and then excreted by the kidneys, its flow rate increases about fivefold as the tubular fluid passes through the proximal tubules; then its rate of flow remains constant in the other tubules. This is because large quantities of PAH are *secreted* into the tubular fluid by the proximal tubular epithelial cells, and it is not reabsorbed in any segment of the tubular system.

These two substances play an important role in experimental studies of tubular function, as will be discussed later in the chapter.

Transport of Different Ions by the Tubular Epithelium—Sodium, Chloride, Bicarbonate, and Potassium. Figure 34–12 illustrates the rates of flow of different important ions—sodium, chloride, bicarbonate, and potassium—in different segments of the tubular system. Note that all these decrease markedly because of reabsorption as the tubular fluid progresses from glomerular filtrate to urine.

In most segments of the tubules, positive ions are generally transported through the tubular epithelium by active transport processes; negative ions are usually transported passively as a result of electrical differences developed across the membrane when the positive ions are transported. For instance, when sodium ions are transported out of the proximal tubular fluid, the resulting electronegativity that develops in the tubular fluid causes chloride ions to follow in the wake of the sodium ions. But, despite the general rule that negative ions are usually passively reabsorbed, this may not be entirely true because some evidence exists for specific active transport of chloride ions in the distal tubule.

Secretion of Potassium and Hydrogen Ions. Note also in Figure 34–12 that the flow of potassium ions normally increases as the tubular fluid passes through the distal tubules and collecting tubules, as a result of *secretion of potassium.* This

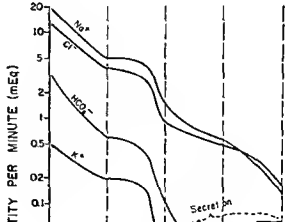


Figure 34–12. Rates of flow of the different ions at different points in the tubules each minute. Note not only the reabsorption of the ions but also the secretion of potassium into the distal and collecting tubules.

will be discussed at greater length in Chapter 36 in relation to the regulation of potassium concentration in the extracellular fluids.

Hydrogen ions are actively secreted in the proximal tubules, distal tubules, and collecting ducts. This secretion is an essential ingredient of the body's system for controlling the acid-base balance of the internal environment (the extracellular fluid), as will be discussed in Chapter 37.

Special Aspects of Bicarbonate Ion Transport. Bicarbonate ion is probably mainly reabsorbed in the form of carbon dioxide rather than in the form of bicarbonate ion itself. This occurs as follows: The bicarbonate ions in the tubular fluid first combine with hydrogen ions that are secreted into the fluid by the epithelial cells. The reaction forms carbonic acid which then dissociates into water and carbon dioxide. The carbon dioxide, being highly lipid-soluble, diffuses rapidly through the tubular membrane into the peritubular capillary blood. When more bicarbonate ions are present than there are available hydrogen ions, most of the excess bicarbonate ions flow on into the urine because the tubules are only slightly permeable to them.

Transport of Other Ions. Though we know much less about the specific means of transport of other ions besides the four illustrated in Figure 34–12, in general essentially all of them can be reabsorbed either by active transport or as a result of electrochemically driven diffusion across the membrane. Thus, calcium, magnesium, and other positive ions are actively reabsorbed, and many of the negative ions are reabsorbed as a result of electrochemical differences that develop when the positive ions are reabsorbed. In addition, certain negative ions—urate, phosphates, sulfate, and nitrate—can be reabsorbed by active transport, this occurring to the greatest extent in the proximal tubules.

CONCENTRATIONS OF DIFFERENT SUBSTANCES AT DIFFERENT POINTS IN THE TUBULES

Whether or not a substance becomes concentrated in the tubular fluid as it moves along the tubules is determined by the *relative reabsorption of the substance versus the reabsorption of water.* If a greater percentage of water is reabsorbed, the substance becomes more concentrated. Conversely, if a greater percentage of the substance is reabsorbed, it becomes more dilute. Therefore, we can combine the curves of Figures 34–10, 34–11, and 34–12 to determine which substances become concentrated in the tubular system and which become diluted. Figure 34–13 illustrates the results of this combination, showing three different classes of [illegible] as follows:

First, the *nutritionally important substances—glucose, protein,* and *amino acids*—are reabsorbed so much more rapidly than water that their concentrations fall extremely rapidly in the proxi[illegible] tubules and remain essentially zero throug[illegible] the remainder of the tubular system as well as the urine.

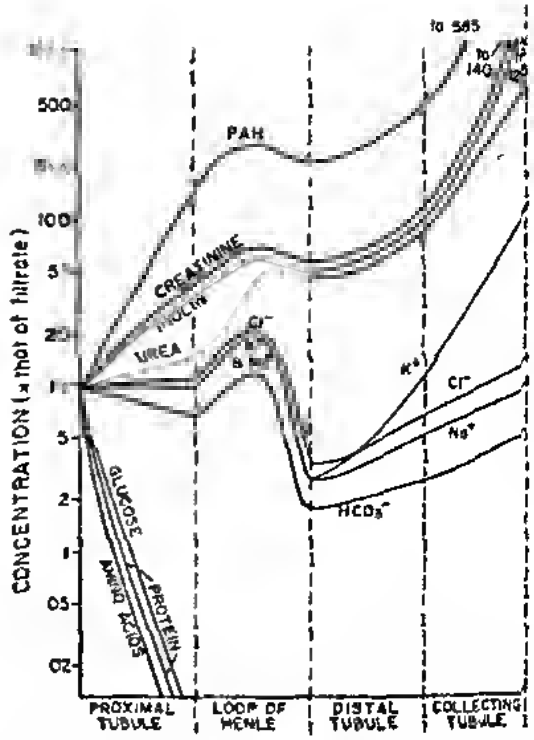


Figure 34–13. Composite figure showing average conce[illegible] tions of different substances at different points in the tu[illegible] system.

Second, the concentrations of the *metabolic e[illegible] products* as well as of the artificially injected s[illegible] stances inulin and para-aminohippuric acid (PA[illegible] become progressively greater throughout the [illegible] bular system, because all these substances [illegible] reabsorbed to a far less extent than is water. N[illegible] that potassium ions also fall into this categ[illegible] because of the secretion of potassium ions into [illegible] distal tubules.

Third, many *other ions are normally excre[illegible] into the urine in concentrations not greatly differ[illegible] from those in the glomerular filtrate and extrac[illegible] lular fluid.* For instance, sodium and chloride io[illegible] on the average, are normally reabsorbed from [illegible] tubules in not too dissimilar proportions fr[illegible] water.

Table 34–1 summarizes the concentrating a[illegible] ity of the tubular system for different substanc[illegible] It also gives the actual quantities of the differ[illegible] substances normally handled by the tubules e[illegible] minute.

Variability of Concentrating Power. A w[illegible] of caution must be given at this point: the qu[illegible] titative values for reabsorption of different s[illegible] stances as given in Figures 34–10, 34–11, and [illegible] 12 and for concentrating power of the tubules given in Figure 34–13 and Table 34–1 are *average* values. These values change greatly [illegible] different physiological states. For instance, w[illegible]

Table 34–1 RELATIVE CONCENTRATIONS OF SUBSTANCES IN THE GLOMERULAR FILTRATE AND IN THE URINE

	Glomerular filtrate (*125 ml/min*)		Urine (*1 ml/min*)		Conc. Urine/Conc. Plasma (*Plasma clearance per min*)
	Quantity/min	Concentration	Quantity/min	Concentration	
Na^+	17 7 mEq	142 mEq/L	0 128 mEq	128 mEq/L	0 9
K^+	0 63	5	0 06	60	12
Ca^{++}	0 5	4	0 0048	4 8	1.2
Mg^{++}	0 38	3	0 015	15	5 0
Cl^-	12.9	103	0 134	134	1.3
HCO_3^-	3 5	28	0 014	14	0 5
$H_2PO_4^-$, HPO_4^{--}	0 25	2	0 05	50	25
SO_4^{--}	0 09	0 7	0 033	33	47
Glucose	125 mg	100 mg/dl	0 mg	0 mg/dl	0 0
Urea	33	26	18 2	1820	70
Uric acid	3 8	3	0 42	42	14
Creatinine	1 4	1 1	1 96	196	140
Inulin	—	—	—	—	125
Diodrast	—	—	—	—	560
PAH	—	—	—	—	585

essentially no antidiuretic hormone is secreted by the hypothalamic-posterior pituitary mechanism, reabsorption of water by the collecting tubules and collecting ducts becomes greatly reduced; and the presence of the extra water in the urine dilutes essentially all the substances listed in Table 34–1 by as much as fourfold. Conversely, secretion of large quantities of antidiuretic hormone causes such excessive reabsorption of water that all the substances can be concentrated to as much as four to five times the values given in Table 34–1.

THE CONCEPT OF "PLASMA CLEARANCE"

The term "plasma clearance" is used to express the ability of the kidneys to clean, or "clear," the plasma of various substances. Thus, if the plasma passing through the kidneys contains 0.1 gram of a substance in each deciliter, and 0.1 gram of this substance also passes into the urine each minute, then 1 deciliter of the plasma is cleaned or cleared of the substance per minute.

Referring again to Table 34–1, note that the normal concentration of urea in each milliliter of plasma and glomerular filtrate is 0 26 mg, and the quantity of urea that passes into the urine each minute is approximately 18.2 mg. Therefore, the equivalent quantity of plasma that completely loses its entire content of urea each minute can be calculated by dividing the quantity of urea entering the urine each minute by the quantity of urea in each milliliter of plasma. Thus 18 2 ÷ 0 26 = 70, that is, 70 ml of plasma is *cleaned* or *cleared* of urea each minute. This amount that is cleared each minute is known as the *plasma clearance* of urea. Obviously, the plasma clearance is a measure of the effectiveness of the kidney to remove substances from the extracellular fluid.

Plasma clearance for any substance can be calculated by the formula below.

$$\text{Plasma clearance (ml/min)} = \frac{\text{Quantity of urine (ml/min)} \times \text{Concentration in urine}}{\text{Concentration in plasma}}$$

The plasma clearances of the usual constituents of urine are shown in the last column of Table 34–1.

INULIN CLEARANCE AS A MEASURE OF GLOMERULAR FILTRATION RATE

Inulin is a polysaccharide that has the specific attributes of not being reabsorbed to a significant extent by the tubules of the nephron and yet being of small enough molecular weight (about 5200) that it passes through the glomerular membrane as freely as the crystalloids and water of the plasma. Also, inulin is not actively secreted even in the minutest amount by the tubules. Consequently, glomerular filtrate contains the same concentration of inulin as does plasma, and as the filtrate flows down the tubules all the filtered inulin continues on into the urine. Thus, *all the glomerular filtrate formed is cleared of inulin.* Therefore, the *plasma clearance per minute of inulin is equal to the glomerular filtration rate.*

As an example, let us assume that it is found by chemical analysis that the inulin concentration in the plasma is 0.1 gram in each 100 ml, and that 0 125 gram of inulin passes into the urine per minute. By dividing 0.1 into 0 125, one finds that 1.25 *100-milliliter portions* of glomerular filtrate must be formed each minute in order to deliver to the urine the analyzed quantity of inulin. In other words, 125 ml of glomerular filtrate is formed per minute, and this is also the plasma clearance of inulin

Inulin is not the only substance that can be used for determining the quantity of glomerular filtrate formed each minute, for the plasma clearance of any other substance that is totally diffusible through the glomerular membrane but is neither absorbed nor secreted by the tubular walls is equal to the glomerular filtration rate. *Mannitol* is a monosaccharide that is frequently used instead of inulin for such measurements, and *radioactive iothalamate* is still another substance frequently used because its radioactivity allows easy quantitative analysis.

PARA-AMINOHIPPURIC ACID (PAH) CLEARANCE AS A MEASURE OF PLASMA FLOW THROUGH THE KIDNEYS

PAH, like inulin, passes through the glomerular membrane with perfect ease. However, it is different from inulin in that most of the PAH remaining in the plasma after the glomerular filtrate is formed is secreted into the tubules by the proximal tubular epithelium if the plasma concentration of PAH is very low. Indeed, only about one tenth of the original PAH remains in the plasma by the time the blood leaves the kidneys.

One can use the clearance of PAH for estimating the *flow of plasma* through the kidneys. As an example, let us assume that 1 mg of PAH is present in each 100 ml of plasma, and that 5.85 mg of PAH passes into the urine per minute. Consequently, 585 ml of plasma is cleared of PAH each minute. Obviously, if this much plasma is cleared of PAH, *at least this much plasma* must have passed through the kidneys in this same period of time. And, since it is known that almost all the PAH is cleared from the blood as it passes through the kidneys, 585 ml would be a reasonable first approximation of the actual plasma flow per minute.

Yet, to be still more accurate, one can correct for the average amount of PAH that is still in the blood when it leaves the kidney. In different experiments the PAH clearance has averaged 91 per cent of the plasma load of PAH entering the kidneys. Therefore, the 585 ml of plasma just calculated would be only 91 per cent of the total amount of plasma flowing through the kidneys. Dividing 585 by 0.91 gives a total plasma flow per minute of approximately 650 ml.

The 91 per cent removal of PAH as it passes through the kidneys is called the *extraction ratio*.

One can calculate the *total blood flow* through the kidneys each minute from the plasma flow and the hematocrit (the percentage of red blood cells in the blood). If the hematocrit is 45 per cent and the plasma flow 650 ml/min, the total blood flow through both kidneys is 650 times (100/55), or 1182 ml/min.

Clearance of *Diodrast* can also be used for estimating plasma flow or blood flow through the two kidneys. However, the extraction ratio of Diodrast averages about 0.85, slightly less than that of PAH.

CALCULATING THE FILTRATION FRACTION FROM PLASMA CLEARANCES

To calculate the filtration fraction—that is, the fraction of the plasma that filters through the glomerular membrane—one must determine (1) the plasma flow through the two kidneys (PAH clearance) and (2) the glomerular filtration rate per minute (inulin clearance). Using 650 ml plasma flow and 125 ml glomerular filtration rate as normal values, we find that the calculated filtration fraction is 125/650, or, to express this as a percentage, 19 per cent.

EFFECT OF "TUBULAR LOAD" AND "TUBULAR TRANSPORT MAXIMUM" IN URINE CONSTITUENTS

Tubular Load. The *tubular load* of a substance is the total amount of the substance that filters through the glomerular membrane into the tubules each minute. For instance, if 125 ml of glomerular filtrate is formed each minute with a glucose concentration of 100 mg per cent, the tubular load of glucose is 100 mg × 1.25, or 125 mg of glucose per minute. Similarly, the load of sodium that enters the tubules each minute is approximately 17 mEq/min, the load of chloride ion is about 13 mEq/min, and the load of urea is approximately 33 mg/min, etc.

Maximum Rate of Transport of Actively Reabsorbed or Secreted Substances—The "Tubular Transport Maximum" (Tm). Since each substance that is actively reabsorbed requires a specific transport system in the tubular epithelial cells, the maximum amount that can be reabsorbed often depends on the maximum rate at which the transport system itself can operate, and this in turn depends on the total amounts of carrier and specific enzymes available. Consequently, for almost every actively reabsorbed substance there is a maximum rate at which it can be reabsorbed, this is called the *tubular transport maximum* for the substance, and the abbreviation is Tm. For instance, the Tm for glucose averages 320 mg/min for the adult human being and if the tubular load becomes greater than 320 mg/min, the excess above this amount is not reabsorbed but instead passes on into the urine. Ordinarily, though, the tubular load of glucose is only 125 mg/min, so that for practical purposes all of it is reabsorbed.

Figure 34–14 demonstrates the relationship between tubular load of glucose, tubular transport maximum for glucose, and rate of glucose loss into the urine. Note that when the tubular load is at its normal level of 125 mg/min there is no detectable loss of glucose into the urine. However, when the tubular load rises above about 220 mg/min, significant quantities of glucose begin to appear in the urine. And, once the load has risen above approximately 400 mg/min, the loss into the urine thereafter is equal to the *tubular load minus 320 mg/min*. Thus at 400 mg/min tubular load, the loss is 80 mg/min, and at 800 mg/min tubular load it is 480 mg/min. In other words, 320 mg/min of the tubular load, which represents the tubular transport maximum for glucose, is reabsorbed, and all the remainder is lost into the urine.

Threshold for Substances That Have a Tubular Transport Maximum. Every substance that has a transfer maximum also has a "threshold" concentration

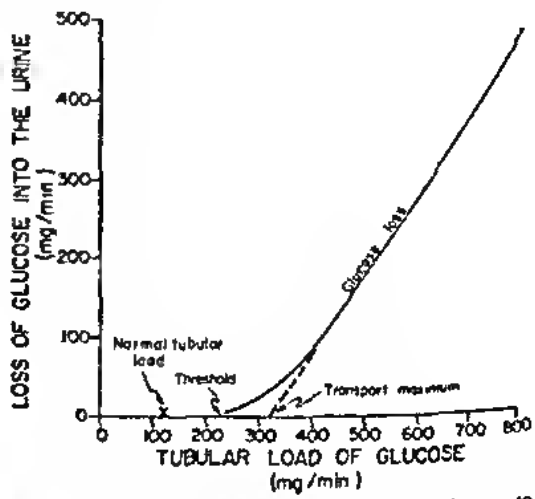


Figure 34–14. Relationship of tubular load of glucose to loss of glucose into the urine.

in the plasma below which none of it appears in the urine and above which progressively larger quantities appear.

Thus, Figure 34–14 shows that glucose begins to spill into the urine when its tubular load exceeds 220 mg/min. The threshold concentration of glucose in plasma that gives this tubular load is 180 mg per cent when the kidneys are operating at their normal glomerular filtration rate of 125 ml/min.

Tubular Transport Maximums of Important Substances Absorbed from the Tubules. Some of the important tubular transport maximums for substances absorbed from the tubules are the following:

Glucose	320	mg/min
Phosphate	0.1	mM/min
Sulfate	0 06	mM/min
Amino acids	1 5	mM/min
Urate	15	mg/min
Plasma protein	30	mg/min
Hemoglobin	1	mg/min
Lactate	75	mg/min
Acetoacetate	variable	(about 30 mg/min)

It is particularly significant that the tubular transport maximum for sodium ions is so large that it has not been measured accurately

Tubular Transport Maximums for Secretion. Substances that are actively secreted by the tubules also exhibit tubular transport maximums as follows

	mg/min
Creatinine	16
PAH	80
Diodrast	57 (of iodine)
Phenol red	56

REFERENCES

Andersson, B Regulation of body fluids *Annu Rev Physiol 39* 185, 1977

Andreoli, T E, and Schaefer, J A Mass transport across cell membranes The effects of antidiuretic hormone on water and solute flows in epithelia *Annu Rev Physiol*, *38* 451, 1976

Andreoli, T, *et al* (eds) Physiology of Membrane Disorders New York, Plenum Press, 1978

Barger, A C, and Herd, J A Renal vascular anatomy and distribution of blood flow *In* Orloff, F, and Berliner, R W (eds) Handbook of Physiology Sec 8 Baltimore, Williams & Wilkins, 1973, p 249

Beeuwkes, R, III The vascular organization of the kidney *Annu. Rev Physiol*, *42* 531, 1980

Berry, C A Heterogeneity of tubular transport processes in the nephron *Annu Rev Physiol*, *44* 181, 1982

Brenner, B M, and Berliner, R W Transport of potassium *In* Orloff, F, and Berliner, R W (eds) Handbook of Physiology Sec 8 Baltimore, Williams & Wilkins, 1973 p 497

Brenner, B M, and Deen, W M The physiological basis of glomerular ultrafiltration *In* MTP International Review of Science Physiology, Vol 6 Baltimore, University Park Press, 1974, p 335

Bulger, R E, and Dobyan, D C Recent advances in renal morphology. *Annu Rev. Physiol*, *44* 147, 1982

Burg, M, and Good, D Sodium chloride coupled transport in mammalian nephrons *Annu Rev Physiol*, *45* 533, 1983

Burg, M, and Stoner, L Renal tubular chloride transport and the mode of action of some diuretics *Annu Rev. Physiol*, *38* 37, 1976

Charlton, C A The Urological System. New York, Churchill Livingstone, 1980

Churg, J, (ed) The Kidney Baltimore, Williams & Wilkins, 1979

Deetjen, P., *et al*, Physiology of the Kidney and of Water Balance New York, Springer Verlag, 1975

Giebisch, G, and Windhager, E Electrolyte transport across renal tubular membranes *In* Orloff, F, and Berliner, R W (eds.) Handbook of Physiology. Sec 8 Baltimore, Williams & Wilkins, 1973, p 315

Glynn, I M, and Karlish, S J D The sodium pump *Annu Rev Physiol*, *37* 13, 1975

Gottschalk, C W, and Lassiter, W E Micropuncture methodology *In* Orloff, F, and Berliner, R W (eds) Handbook of Physiology Sec. 8 Baltimore, Williams & Wilkins, 1973, p 129

Gottschalk, C W, *et al* Micropuncture study of composition of loop of Henle fluid in desert rodents *Am J Physiol*, *204* 532, 1963

Grantham, J. J, *et al* Studies of isolated renal tubules in vitro *Annu. Rev. Physiol*, *40* 249, 1978

Greger, R, *et al*. Renal excretion of purine metabolites, urate and allantoin by the mammalian kidney *Int Rev Physiol*, *11* 257, 1976

Guyton, A C, *et al* Dynamics and Control of the Body Fluids Philadelphia, W B Saunders Co, 1975

Katz, A I, and Lindheimer, M D Actions of hormones on the kidney *Annu Rev Physiol*, *39* 97, 1977

Kinne, R Membrane-molecular aspects of tubular transport. *Int Rev Physiol*, *11* 169, 1976

Knox, F G, and Davis, B B Role of physical and neuroendocrine factors on proximal electrolyte reabsorption *Metabolism*, *23* 793, 1974

Knox, F G, and Spielman, W S Renal circulation *In* Shepherd, J T, and Abboud, F M (eds) Handbook of Physiology Sec 2, Vol III Bethesda, American Physiological Society, 1983, p 183

Koeppen B M, *et al* Electrophysiology of mammalian renal tubules Inferences from intracellular microelectrode studies *Annu Rev Physiol*, *45* 497, 1983

Lameire, N H, *et al* Heterogeneity of nephron function. *Annu Rev Physiol*, *39* 159, 1977

Lassiter, W E Kidney *Annu Rev Physiol* *37* 371, 1975

Latta, H Ultrastructure of the glomerulus and juxtaglomerular apparatus. *In* Orloff, F, and Berliner, R W (eds) Handbook of Physiology Sec 8 Baltimore, Williams & Wilkins, 1973, p 1

Maude, D L Mechanism of tubular transport of salt and water *In* MTP International Review of Science Physiology Vol 6 Baltimore, University Park Press, 1974, p 39

Maunsbach, A B Cellular mechanisms of tubular protein transport *Int Rev Physiol*, *11* 145, 1978

Morel, F, *et al* Distribution of hormone-dependent adenylate cyclase in the nephron and its physiological significance *Annu Rev. Physiol*, *43* 569, 1981

Mudge, G H, *et al* Tubular transport of urea, glucose, phosphate, uric acid, sulfate, and thiosulfate *In* Orloff F, and Berliner, R W (eds) Handbook of Physiology Sec 8. Baltimore, Williams & Wilkins, 1973, p 587

Pitts, R. Physiology of the Kidney and Body Fluids 3rd Ed Chicago, Year Book Medical Publishers, 1974

Renkin, E M, and Gilmore, J P Glomerular filtration *In* Orloff, F, and Berliner, R W (eds) Handbook of Physiology Sec 8 Baltimore, Williams & Wilkins, 1973, p 185

Ross, C R, and Holohan, P D Transport of organic anions and cations in isolated renal plasma membranes *Annu Rev Pharmacol. Toxicol*, *23* 65, 1983

Smith, H W. The Kidney Structure and Function in Health and Disease New York, Oxford University Press, 1951.

Stoff, J S, *et al* Recent advances in renal tubular biochemistry *Annu Rev Physiol*, *38* 46, 1976

Ullrich, K L Sugar, amino acids, and Na+ cotransport in the proximal tubule *Annu Rev Physiol*, *41* 181, 1979.

Vander, A J: Renal Physiology New York, McGraw Hill, 1980.

Walker, L A, and Valtin, H Biological importance of nephron heterogeneity *Annu Rev. Physiol*, *44* 203, 1982

Windhager, E E Kidney, water, and electrolytes *Annu Rev. Physiol*, *31* 117, 1969

Wright, F. S: Potassium transport by the renal tubule *In* MTP International Review of Science Physiology. Vol 6. Baltimore, University Park Press, 1974, p 79.

35

Renal Mechanisms for (1) Controlling Glomerular Filtration, (2) Excreting Dilute or Concentrated Urine, and (3) Excreting Urea, Sodium, Potassium, and Fluid Volume

In the previous chapter we discussed the mechanisms by which glomerular filtrate is formed and the way in which it is then processed in the tubules to become urine. The present chapter will be devoted to special mechanisms that function within the kidney to control the composition of urine from moment to moment to reflect the needs of the body for excreting different substances. For instance, the body at times has great excesses of water and, therefore, must excrete a dilute urine; at other times it has a deficit of water and, therefore, must excrete a minimum of water, yet still excrete the end-products of metabolism as well as other substances that must be removed from the body fluids.

Specifically, we will discuss in this chapter (1) the special mechanisms for controlling the glomerular filtration rate, which is a prerequisite for essentially all other control mechanisms of the kidney to function properly; (2) the special renal mechanisms that allow the kidneys to excrete either a dilute or a concentrated urine; and (3) the tubular mechanisms that allow the kidneys to excrete appropriate quantities of urea, sodium, potassium, and fluid volume.

AUTOREGULATION OF GLOMERULAR FILTRATION RATE

The glomerular filtration rate normally remains very nearly constant hour after hour, usually varying very little either above or below the norm value of about 125 ml per minute for the t kidneys Even a change in arterial pressure fr as little as 75 mm Hg to as high as 160 mm hardly changes the glomerular filtration rate. Tl effect is illustrated in Figure 35-1, and it is call *autoregulation of glomerular filtration rate.*

Why Is It Important for the Glomerular F tration Rate to Be Autoregulated? To und stand the importance of maintaining a consta glomerular filtration rate, let us consider wh would happen if the glomerular filtration ra should become, first, very slight or, second, ve great.

At a very slight glomerular filtration rate, tl tubular fluid would pass through the tubules slowly that essentially all of it would be rea sorbed. Therefore, the kidney would fail to elim nate the necessary waste products.

At the other extreme, with a much too hig glomerular filtration rate, the fluid would pass s rapidly through the tubules that they would b unable to reabsorb those substances that need t be conserved in the body.

Thus, one can readily see that the glomerula filtrate roust flow into the tubular system at a appropriate rate (1) to allow the unwanted sub stances to pass on into the urine while (2) reab sorbing the wanted substances. However, it is ofte not appreciated how narrow the range of glomer

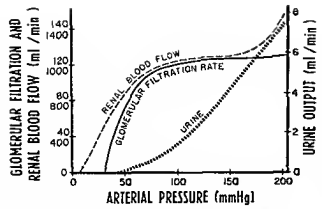


Figure 35–1. Autoregulation of glomerular filtration rate (GFR) and renal blood flow (RBF) when the arterial pressure is increased, but there is lack of autoregulation of urine flow.

ular filtration rate must be if optimal function of the tubular system is to occur. To emphasize this narrow range, recent analyses of tubular function have shown that even a 5 per cent too great or too little rate of glomerular filtration can have considerable effects in causing either excess loss of solutes and water into the urine or, at the other extreme, too little of the necessary excretion of waste products.

MECHANISM OF AUTOREGULATION OF GLOMERULAR FILTRATION RATE—TUBULOGLOMERULAR FEEDBACK

The precision with which glomerular filtration rate must be autoregulated demands that there be an effective means for controlling this filtration rate. Fortunately, each nephron is provided with not one but *two* special feedback mechanisms which add together to provide the degree of glomerular filtration autoregulation that is required. These two mechanisms are (1) an *afferent arteriolar vasodilator feedback mechanism*, and (2) an *efferent arteriolar vasoconstrictor feedback mechanism*. In each of these mechanisms, the degree to which the fluid entering the distal tubule has been processed is detected, and appropriate signals are fed back to the afferent and efferent arterioles to increase or decrease the glomerular filtration rate as required. The combination of these two feedback mechanisms is called *tubuloglomerular feedback*. And the feedback process probably occurs either entirely or almost entirely at the *juxtaglomerular complex*, which has the following characteristics:

The Juxtaglomerular Complex. Figure 35–2 illustrates the juxtaglomerular complex, showing that the initial portion of the distal tubule, immediately after the upper end of the thick segment of the ascending limb of the loop of Henle, passes in the angle between the afferent and efferent arterioles, actually abutting each of these two arterioles. Furthermore, those epithelial cells of the distal tubule that come in contact with the arterioles are more dense than the other tubular cells and are collectively called the *macula densa*. The macula densa cells appear to secrete some substance toward the arterioles because the Golgi apparatus, an intracellular secretory organelle, is directed toward the arterioles and not toward the lumen of the tubule, in contrast to all the other

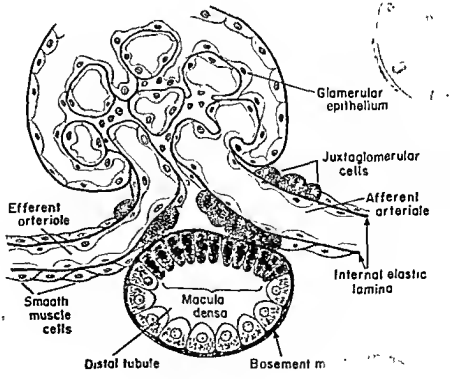


Figure 35–2. Structure of the juxtaglomerular apparatus, illustrating its possible feedback role in the control of nephron function. (Modified from Ham, Histology. J. B. Lippincott Co.)

tubular epithelial cells. Note also in Figure 35–2 that the smooth muscle cells of both the afferent and efferent arterioles are swollen and contain dark granules where they come in contact with the macula densa. These cells are called *juxtaglomerular cells*, and the granules are composed mainly of inactive *renin*. The whole complex of macula densa and juxtaglomerular cells is called the *juxtaglomerular complex*.

Thus, the anatomical structure of the juxtaglomerular apparatus suggests strongly that the fluid in the distal tubule in some way plays an important role in helping to control nephron function by providing feedback signals to both the afferent and efferent arterioles.

The Afferent Arteriolar Vasodilator Feedback Mechanism

A low rate of glomerular filtration causes overreabsorption of sodium and chloride ions in the ascending limb of the loop of Henle and therefore decreases the ionic concentration at the macula densa. This decrease in ions in turn initiates a signal from the macula densa to dilate the afferent arteriole. Putting these two facts together, the following is the postulated mechanism by which the afferent arteriolar vasodilator feedback mechanism controls glomerular filtration rate:

(1) Too little flow of glomerular filtrate into the tubules causes decreased sodium and chloride ion concentration at the macula densa.

(2) The decreased ionic concentration causes afferent arteriolar dilatation.

(3) This in turn allows increased blood flow into the glomerulus, which increases the glomerular pressure.

(4) The increased glomerular pressure, as well as the increased glomerular blood flow, increases the glomerular filtration rate back toward the required level.

Thus, this is a typical negative feedback mechanism for controlling the glomerular filtration rate at a steady rate. This mechanism also helps autoregulate the renal blood flow at the same time, as we shall discuss.

The Efferent Arteriolar Vasoconstrictor Feedback Mechanism

Too few sodium and chloride ions at the macula densa are believed also to cause the juxtaglomerular cells to release renin, and this in turn causes formation of angiotensin. The angiotensin then constricts mainly the efferent arteriole because it is highly sensitive to angiotensin II, much more so than is the afferent arteriole.

With these facts in mind, we can now describe the efferent arteriolar vasoconstrictor mechanism that helps maintain a constant glomerular filtration rate:

(1) A too low glomerular filtration rate causes excess reabsorption of sodium and chloride ions in the ascending limb of the loop of Henle, reducing the ionic concentration at the macula densa.

(2) The low concentration of ions then causes the JG cells to release renin from their granules.

(3) The renin causes formation of angiotensin II.

(4) The angiotensin II constricts the efferent arterioles, which causes the pressure in the glomerulus to rise.

(5) The increased pressure then causes the glomerular filtration rate to return back toward normal.

Thus, this is still another negative feedback mechanism that helps to maintain a very constant glomerular filtration rate; it does so by constricting the efferent arterioles at the same time that the afferent vasodilator mechanism described above dilates the afferent arterioles. When both these mechanisms function together, the glomerular filtration rate increases only a few per cent even though the arterial pressure changes between the limits of 75 and 160 mm Hg.

Role of the Renin-Angiotensin System and the Efferent Vasoconstrictor Mechanism in Conserving Water and Salt but Eliminating Waste Products During Arterial Hypotension. When the arterial pressure falls too low, the intrarenal renin-angiotensin system not only helps maintain normal glomerular filtration (by constricting the efferent arterioles), but it also provides a means for conserving as much water and salt as possible to keep the arterial pressure from falling any lower while at the same time allowing the kidneys to continue excreting the body's waste products. Let us explain this.

As we shall learn later in this chapter, the amount of urea and other waste products excreted by the kidneys is determined to a great extent by the rate of glomerular filtration. The reason for this is that once the waste products have been filtered through the glomeruli into the tubules, as much as half of these are not reabsorbed by the tubules and therefore pass on into the urine. Therefore, the fact that angiotensin keeps the glomerular filtration rate near normal also keeps the rate of waste product excretion almost normal as well.

On the other hand, angiotensin II has several effects that cause increased water and salt reabsorption by the tubules. These are:

(1) When the angiotensin constricts the efferent arterioles, this reduces the renal blood flow, which then reduces the peritubular capillary pressure. This increases the rate of reabsorption of water and electrolytes from the tubular system, especially from the proximal tubules because the balance of forces at the capillary membrane is now in favor of absorption.

(2) Because the blood flow through the glomerulus is now decreased and yet the rate of glomerular filtration is still near normal, a very high proportion of the fluid of the plasma filters through the glomerulus into the tubules. Therefore, the concentration of the plasma proteins in the blood leaving the glomeruli becomes very high, and this concentrated plasma flows on into the peritubular capillaries. As a result, the colloid osmotic pressure in the peritubular capillaries is greatly increased, which is an additional factor that enhances reabsorption of water and salt.

3) Some research workers have suggested that angiotensin also has a direct effect on the distal tubules in causing increased active reabsorption of sodium. If true, this could be still another factor that would increase the absorption of salt and water.

Regardless of which mechanisms are involved, the action of the renin-angiotensin system is to cause conservation of water and electrolytes while at the same time promoting continued high excretion of waste products. Obviously, this is an exceedingly important life-saving mechanism when arterial hypotension occurs. This mechanism is also important when animals or human beings become dehydrated, such as when they are exposed to desert conditions, because dehydration alone also increases the intrarenal renin secretion to conserve water and salt while allowing the waste products to be excreted.

AUTOREGULATION OF RENAL BLOOD FLOW

Not only is the glomerular filtration rate autoregulated at a very constant level, but the renal blood flow also remains near the normal level despite considerable change in the arterial pressure, this is called *autoregulation of renal blood flow*. For instance, the renal blood flow of both kidneys remains very near the normal level of 1200 ml per minute even when the arterial pressure falls as low as 75 mm Hg or rises to as high as 160 mm Hg. This, too, is illustrated in Figure 35–1.

The most important mechanism for autoregulating the renal blood flow is probably the first of the two mechanisms that autoregulates glomerular filtration rate, the *afferent arteriolar vasodilator mechanism*. That is, when the renal blood flow falls too low, this at first decreases the glomerular filtration rate. The falling glomerular filtration rate in turn causes a feedback effect at the juxtaglomerular complex to dilate the afferent arteriole, as explained earlier. This then allows increased blood flow into the glomerulus, thus returning both the renal blood flow and the glomerular filtration rate back toward normal. It should be recognized, however, that this mechanism is primarily concerned with autoregulating glomerular filtration rate, which is a necessary condition for normal function of the nephron. The autoregulation of renal blood flow is merely incidental to the autoregulation of the filtration.

Some physiologists believe that still another mechanism is important in autoregulating the renal blood flow. This is the so-called *myogenic mechanism*, which has been postulated as follows. When the arterial pressure rises, it stretches the wall of the arteriole, and this in turn causes a secondary contraction of the arteriole. This decreases the renal blood flow back toward normal, thus opposing the effect of the rising arterial pressure to increase the flow. Conversely, when the pressure falls too low, an opposite myogenic response allows the artery to dilate and therefore increases the flow. Unfortunately, this mechanism has not yet been proved, and there are many reasons for believing that it would be inherently unstable, so that some physiologists do not believe that such a mechanism truly does help in autoregulating the renal blood flow. Obviously, though, if this myogenic mechanism for renal blood flow regulation should prove to be a real phenomenon, it could help autoregulate the glomerular filtration rate as well as the renal blood flow.

THE MECHANISM FOR EXCRETING EXCESS WATER: EXCRETION OF A DILUTE URINE

One of the most important functions of the kidney is to control the osmolality of the body fluids. When the osmolality falls too low—that is, when the fluids become too dilute—the kidneys automatically excrete a great excess of water in the urine. This obviously causes a dilute urine, but it also removes water from the body, in this way increasing the body fluid osmolality back toward normal. Conversely, when the osmolality of the body fluids is too great, the kidneys automatically excrete an excess of solutes, thereby reducing the body fluid osmolality again back toward normal, but at the same time excreting a concentrated urine.

Role of Antidiuretic Hormone in Controlling Urine Concentration. The signal that tells the kidney whether to excrete a dilute or a concentrated urine is the hormone called *antidiuretic hormone* (also called "vasopressin") that is secreted by the posterior pituitary gland. When the body fluids are too concentrated, the posterior pituitary gland secretes large amounts of antidiuretic hormone, which causes the kidneys to excrete excessive amounts of solutes but to conserve water in the body. Conversely, in the absence of antidiuretic hormone, the kidneys excrete a dilute urine, thus removing excessive amounts of water from the body fluids and causing them to become concentrated once again. The feedback mechanism of this system will be described in the following chapter; for now, let us discuss the renal mechanisms for excreting a dilute or a concentrated urine.

The Renal Mechanism for Excreting a Dilute Urine. Figure 35–3 illustrates the mechanism for excreting a dilute urine. When the glomerular filtrate is initially formed by the glomerulus its osmolality is almost exactly the same as that of the plasma, approximately 300 milliosmoles per liter. To excrete excess water, it is necessary to dilute the filtrate as it passes through the tubules. This is achieved by reabsorbing a higher proportion of solutes than water.

The colored arrows in Figure 35–3 represent reabsorption of most of the tubular solutes (besides the waste products) in all the distal segments of the tubular system. These same tubular segments are depicted in the figure with thickened, darkened walls to indicate that these portions are almost completely impermeable to water when the kidneys are excreting a dilute urine. The ascending limb of the loop of Henle and the diluting segment of the distal tubule even normally are very impermeable to water, while the late distal tubule, the cortical collecting tubule, and the collecting ducts also become almost completely impermeable to water *when antidiuretic hormone is not present* in the circulating body fluids.

The reabsorption of solutes in these distal seg-

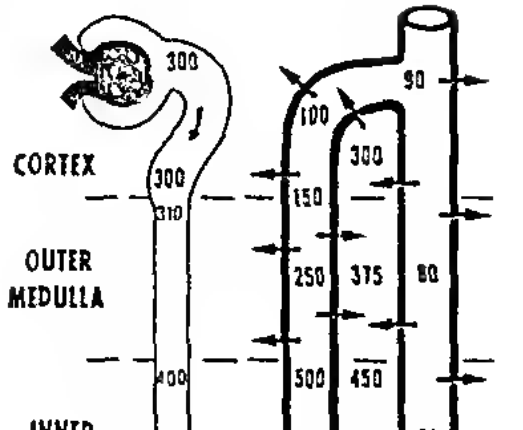


Figure 35–3. The renal mechanism for forming a dilute urine. The darkened walls of the distal portions of the tubular system indicate that these portions of the tubules are relatively impermeable to the reabsorption of water in the absence of antidiuretic hormone. The solid arrows indicate active processes for absorption of most of the solutes besides the urinary waste products. (Numerical values are in milliosmoles per liter.)

ments of the tubular system is strong and active. In the thick segment of the ascending limb of the loop of Henle active reabsorption of sodium, potassium, and chloride ions is especially powerful, and one can see from the numerical values in the figure that the osmolality of the fluid in the ascending limb of the loop of Henle decreases progressively to about 100 milliosmoles per liter by the time the fluid leaves this tubular segment. That is, most of the solutes besides the waste products are reabsorbed, but the water remains behind. Then, as this remaining dilute fluid passes through the distal tubule, cortical collecting tubule, and collecting duct, some additional reabsorption of solutes, especially of sodium ions, causes still more dilution of the tubular fluid, sometimes decreasing its osmolality to as little as 65 to 70 milliosmoles per liter by the time it leaves the collecting duct to enter the urine.

To summarize, the process for excreting a dilute urine is very simple, it is one of absorbing solutes from the distal segments of the tubular system while leaving the water in the tubules. However, this failure of water to be reabsorbed occurs only in the absence of antidiuretic hormone.

THE MECHANISM FOR EXCRETING EXCESS SOLUTES: THE COUNTERCURRENT MECHANISM FOR EXCRETING A CONCENTRATED URINE

The process for concentrating urine is not nearly so simple as for diluting it. Yet, at times it is exceedingly important to concentrate the urine as much as possible so that excess solutes can be eliminated with as little loss of water from the body as possible—for instance, when one is exposed to desert conditions with an inadequate supply of water. Fortunately, the kidneys have developed a special mechanism for concentrating the urine called the *countercurrent mechanism*.

The countercurrent mechanism depends on a peculiar anatomical arrangement of the loops of Henle and the vasa recta. In the human being, the loops of Henle of one third to one fifth of the nephrons dip deep into the medulla and then return to the cortex; some dip all the way to the tips of the papillae that project into the renal pelvis. This group of nephrons with the long loops of Henle is called the *juxtamedullary nephrons*. Paralleling the long loops of Henle are loops of peritubular capillaries called the *vasa recta*; these also loop down into the medulla from the cortex and then back out to the cortex. These arrangements of the different parts of the juxtamedullary nephron and the vasa recta are diagrammed in Figure 35–4.

Hyperosmolality of the Medullary Interstitial Fluid, and Mechanisms for Achieving It

The first step in the excretion of excess solutes in the urine—that is, for excretion of a concentrated urine—is to create hyperosmolality of the medullary interstitial fluid. As we shall see later, this in turn is necessary for concentrating the urine. But, first, let us explain the mechanism for creating this hyperosmolality in the medullary interstitium.

The normal osmolality of the fluids in almost all parts of the body is about 300 mOsm/liter. However, as shown by the numbers in Figure 35–4, the osmolality of the interstitial fluid in the medulla of the kidney is much higher than this, and it becomes progressively greater the deeper one goes into the medulla, increasing from 300 mOsm/liter in the cortex to 1200 mOsm/liter (occasionally as high as 1400 mOsm/liter) in the pelvic tip of the medulla. Four different solute-concentrating mechanisms are responsible for this hyperosmolality; these are:

First, the principal cause of the greatly increased medullary osmolality is active transport of sodium ions (plus cotransport of potassium and chloride ions) out of the thick portion of the ascending limb of the loop of Henle. The large colored arrows shown in this tubular segment in Figure 35–4 illustrate this transport especially of sodium and chloride ions (as well as potassium, calcium, and magnesium ions to a lesser extent) out of the thick portion of the loop of Henle and into the outer medullary interstitial fluid. All these solutes become concentrated in this fluid. These ions are also carried downward into the inner medulla by the

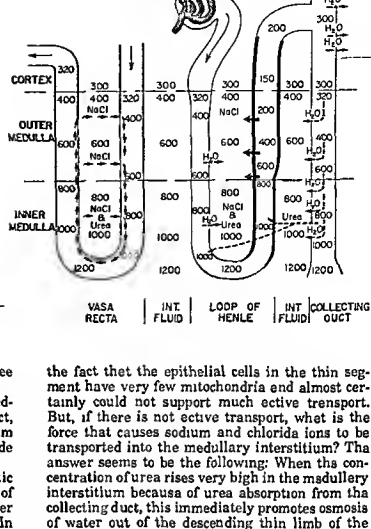


Figure 35–4. The countercurrent mechanism for concentrating the urine (Numerical values are in milliosmoles per liter)

flowing blood in the vasa recta, as we shall see shortly.

Second, ions are also transported into the medullary interstitial fluid from the collecting duct, mainly resulting from active transport of sodium ions and electrogenic passive absorption of chloride ions along with the sodium ions.

Third, when the concentration of antidiuretic hormone is high in the blood, large amounts of urea are also absorbed into the fluid of the inner medulla from the collecting duct. The reason is: In the presence of antidiuretic hormone, the inner medullary portion of the collecting duct becomes moderately permeable to urea and highly permeable to water. The fact that it becomes highly permeable to water causes rapid reabsorption of water out of the collecting duct, and this greatly increases the concentration of urea in the duct. Now, because of this high urea concentration, the urea too diffuses through the collecting duct wall into the medullary interstitium. Consequently, the urea concentration in the medullary interstitial fluid rises almost to equal the concentration in the collecting duct. In the human being, during maximal antidiuretic hormone stimulation this may be to as high as 400 to 500 mOsm/liter, which obviously greatly increases the osmolality of the inner medullary interstitial fluid.

And, fourth, the final event that causes increased osmolal concentration of the medullary interstitial fluid is transport of sodium and chloride ions into the inner medullary interstitium from the thin segment of the loop of Henle. Some physiologists believe that this results from active transport from the ascending portion of this thin segment. However, most recent experiments suggest that this is very unlikely, which also fits with the fact that the epithelial cells in the thin segment have very few mitochondria and almost certainly could not support much active transport. But, if there is not active transport, what is the force that causes sodium and chloride ions to be transported into the medullary interstitium? The answer seems to be the following: When the concentration of urea rises very high in the medullary interstitium because of urea absorption from the collecting duct, this immediately promotes osmosis of water out of the descending thin limb of the loop of Henle. Therefore, the concentration of sodium chloride inside the thin limb rises almost to twice normal. And now, because of this high concentration, both sodium and chloride ions diffuse passively out of the thin segment into the interstitium.

In summary, at least four different factors contribute to the marked increase in osmolality in the medullary interstitial fluid. These are (1) active transport of the ions into the interstitium by the thick portion of the ascending limb of the loop of Henle, (2) active transport of ions from the collecting duct into the interstitium, (3) passive diffusion of large amounts of urea from the collecting duct into the interstitium, and (4) transport of additional sodium and chloride into the interstitium from the thin segment of the loop of Henle, a transport that is probably also passive as explained above. The net result is an increase in the osmolality of the medullary interstitial fluid, when adequate amounts of antidiuretic hormone are present, to as high as 1200 to 1400 mOsm/liter near the tips of the papillae.

Mechanism by Which Antidiuretic Hormone Increases Water Reabsorption. The precise mechanism by which antidiuretic h

creases water reabsorption by the late distal tubules, collecting tubules, and collecting ducts is not known. However, several established facts about the mechanism are the following: Antidiuretic hormone itself acts not on the luminal membrane of the tubular epithelial cells but instead on the basolateral membrane of these cells. It activates the enzyme *adenyl cyclase* in this membrane, which then causes formation of *cyclic adenosine monophosphate* (cyclic AMP) in the cell cytoplasm. The cyclic AMP then diffuses to the luminal membrane of the cells where it greatly increases the permeability to water. The exact way in which this happens is not known, but electron micrographs have shown increased numbers of microtubules in the vicinity of this luminal membrane after antidiuretic hormone stimulation. Regardless of the mechanism, it is the increase in permeability of the luminal membrane that is responsible for the increased water reabsorption by the late distal tubules, collecting tubules, and collecting ducts.

Countercurrent Exchange Mechanism in the Vasa Recta

We have now discussed the mechanisms by which high concentrations of solutes are achieved in the medullary interstitium. However, without a special medullary vascular system as well, the flow of blood through the interstitium would rapidly remove the excess solutes and keep the concentration from rising very high. Fortunately, the medullary blood flow has two characteristics, both exceedingly important, for maintaining the high solute concentration in the medullary interstitial fluids. These are:

First, the medullary blood flow is very slight in quantity, amounting to only 1 to 2 per cent of the total blood flow of the kidney. Because of this very sluggish blood flow, removal of solutes is minimized.

Second, the vasa recta functions as a *countercurrent exchanger* that also minimizes the washout of solutes from the medulla. This can be explained as follows: A countercurrent fluid exchange mechanism is one in which fluid flows through a long U-tube, with the two arms of the U lying in close proximity to each other so that fluid and solutes can exchange readily between the two arms. This obviously also requires that each of the two arms of the U be highly permeable, which is true for the vasa recta. When the fluids and solutes in the two parallel streams of flow can exchange rapidly, tremendous concentrations of solute can be maintained at the tip of the loop with negligible washout of solute.

Thus, in Figure 35–4 as blood flows down the descending limbs of the vasa recta, sodium chloride and urea diffuse into the blood from the interstitial fluid while water diffuses outward into the interstitium, and these two effects cause the osmolal concentration in the capillary blood to rise progressively higher, to a maximum concentration of 1200 mOsm/liter at the tips of the vasa recta. Then, as the blood flows back up the ascending limbs, the extreme diffusibility of all molecules through the capillary membrane allows essentially all the extra sodium chloride and urea to diffuse back out of the blood into the interstitial fluid while water diffuses back into the blood. Therefore by the time the blood finally leaves the medulla its osmolal concentration is only slightly greater than that of the blood that had initially entered the vasa recta. As a result, blood flowing through the vasa recta carries only a minute amount of the medullary interstitial solutes away from the medulla.

Possibility That the Loop of Henle Functions as a Countercurrent "Multiplier" for Increasing Medullary Sodium Concentration. Some physiologists believe that the combination of the countercurrent fluid exchange mechanism of the vasa recta and the sodium chloride absorption by the thick segment of the loop of Henle functions as a "multiplier" system accounting for the very high sodium chloride concentration in the medulla. The presumed function is: When sodium chloride is reabsorbed from the thick segment into the interstitium of the outer medulla, much of this sodium chloride diffuses into the blood in the vasa recta and is carried downward to the inner medulla. Here much of the sodium chloride diffuses back into the thin segment of the loop of Henle and flows once again in the tubular fluid upward to the thick segment of the loop. Again the sodium chloride is absorbed from the tubular fluid, entering once again the medullary interstitium. Thus by being reabsorbed again and again from the loop of Henle and at the same time being conserved in the medullary interstitium by the vasa recta countercurrent exchange mechanism, the concentration of sodium chloride is believed to be progressively "multiplied". Therefore this is called the *countercurrent multiplier mechanism*.

Mechanism for Excreting a Concentrated Urine—Role of Antidiuretic Hormone

Now that we have explained how the kidney creates hyperosmolality in the medullary interstitium, it becomes a simple matter to explain the mechanism for excreting a concentrated urine thus causing loss of excess solutes from the body fluids while at the same time retaining as much water as possible.

When the concentration of *antidiuretic hormone* in the blood is high, the epithelium of the cortical collecting tubule, the collecting duct, and in some species of animals the late distal tubule as well becomes highly permeable to water. This is illustrated in Figure 35–4 by the thin walls of these segments of the tubular system. Most importantly,

as the tubular fluid flows through the collecting duct, water is pulled by osmosis into the highly concentrated fluid of the medullary interstitium. Thus, the collecting duct fluid also becomes highly concentrated, and it issues from the papilla into the pelvis of the kidney at a concentration of about 1200 mOsm/liter, almost exactly equal to the osmolal concentration of the solutes in the medullary interstitium near the papilla.

Summary of the Osmolal Concentration Changes in the Different Segments of the Tubules

Figure 35–5 illustrates the changes in osmolality of tubular fluid as it passes through the different segments of the tubules. In the proximal tubules, the tubular membranes are so highly permeable to water that whenever a solute is transported across the membrane an almost exact proportionate amount of water crosses the membrane by osmosis at the same time; therefore, the osmolality of the fluid remains almost exactly equal to that of the glomerular filtrate, 300 mOsm/liter, throughout the entire extent of the proximal tubule.

Next, in the loop of Henle, the osmolality rises rapidly because of the countercurrent mechanism explained in the previous few paragraphs. When a concentrated urine is being formed—that is, when the blood concentration of ADH is high—the loop of Henle osmolality rises much higher than when a dilute urine is being formed, because of the large quantity of urea that is passively reabsorbed into the medullary interstitium from the collecting ducts.

Next, beginning in the thick portion of the ascending limb of the loop of Henle, the osmolality falls once again to a very low level, usually to about 100 mOsm/liter.

Finally, in the late distal tubule, cortical collecting tubule, and collecting duct the osmolality depends entirely on the presence or absence of

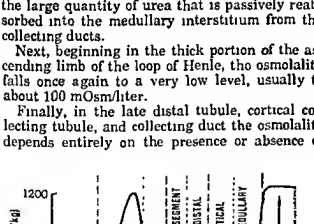


Figure 35–5. Changes in osmolality of the tubular fluid as it passes through the tubular system.

ADH (antidiuretic hormone). In the absence of ADH, very little water is reabsorbed from these segments. Therefore, the osmolality remains less than 100 mOsm/liter and even falls still another few milliosmoles because of active transport of ions through the epithelium of these tubules. Therefore, a very dilute urine is formed.

But, in the presence of excess ADH, the cortical collecting tubule, the collecting duct, and in some animal species the late distal tubule as well become highly permeable to water so that most of the water is reabsorbed, thus producing a very concentrated urine.

Note especially the shaded areas in Figure 35–5. These denote the ranges of concentration of the tubular fluid, and also the range of urine concentration between the extremes of 65 and 1200 mOsm/liter in human beings, depending on the blood concentration of antidiuretic hormone at any given time. The overall mechanism for control of antidiuretic hormone concentration, as well as control of osmolality of the body fluids, is discussed in the following chapter.

OSMOLAR CLEARANCE; FREE WATER CLEARANCE

One can calculate the clearance of osmolar substances in terms of the volume of plasma cleared per minute in the same way that one can calculate the clearance of any single substance. That is, osmolar clearance (C_{osm}) can be calculated by means of the following formula:

$$C_{osm} = \frac{\text{Osmoles entering urine per minute}}{\text{Plasma osmolar concentration}}$$

For instance, if the plasma osmolality is 300 mOsm/kg and the amount of milliosmoles entering the urine per minute is 1.5, then the osmolar clearance will be 1.5/300 liters/min, or 5 ml/min.

Free Water Clearance. When the kidney forms urine that is osmotically more dilute than plasma, it is obvious that a higher proportion of the water in the glomerular filtrate is excreted than of the osmolar substances. The excess water that is excreted is called *free water*, and the total plasma volume that is cleared of this excess water each minute is called the *free water clearance*. The free water clearance can be calculated by first determining the osmolar clearance and then subtracting this rate from the rate of urine flow per minute. Thus, the formula for free water clearance (C_{H_2O}) is:

$$C_{H_2O} = \text{Urine volume per minute} - C_{osm}.$$

The free water clearance is important because it tells one how rapidly the kidneys are changing the body fluid osmolality. The free water clearance can be *positive*, in which case excess water is being removed; or it can be *negative*, in which case excess solutes are being removed.

UREA EXCRETION

The body forms an average of 25 to 30 grams of urea each day—more than this in persons who eat

a very high protein diet and less in persons who are on a low protein diet. All this urea must be excreted in the urine; otherwise, it will accumulate in the body fluids. Its normal concentration in plasma is approximately 26 mg/100 ml, but patients with renal insufficiency frequently have levels as high as 200 mg/100 ml, and it has been recorded as high as 800 mg/dl.

The two major factors that determine the rate of urea excretion are (1) the concentration of urea in the plasma, and (2) the glomerular filtration rate. These factors increase urea excretion mainly because the "load" of urea entering the proximal tubules is equal to the *product of the plasma urea concentration and the glomerular filtration rate.* And, in general, the quantity of urea that passes on through the tubules into the urine averages between 40 and 60 per cent of the urea load that enters the proximal tubules. Only when the function of the kidneys is abnormal does this fraction of the tubular load, called the *tubular fraction* for urea excretion, change much from this normal range.

Effect of Decreased Glomerular Filtration Rate on Urea Excretion and on Plasma Urea Concentration. In many renal diseases, the glomerular filtration rate of the two kidneys falls considerably below normal. And, because urea excretion is a function of glomerular filtration of urea, this means decreased excretion of urea when the glomerular filtration decreases. However, the body continues to form large quantities of urea, which means that urea will progressively collect in the body fluids until its plasma concentration rises very high. Then, the quantity of urea filtered in the glomerular filtrate, which is equal to the plasma urea concentration times the glomerular filtration rate, eventually will become great enough to allow excretion of the urea as rapidly as it is formed. Yet, one will recognize that this occurs only because the plasma concentration of urea has risen very high, and this in itself is a very abnormal state that can be severely damaging to the body.

Therefore, one of the most important reasons for forming large amounts of glomerular filtrate in the two kidneys is to excrete the necessary amounts of urea. Many of the other waste products that must be excreted by the kidneys obey the same principles for excretion as urea, for their rates of excretion also depend highly upon the amount of glomerular filtrate formed by the kidneys. Such substances include creatinine, uric acid, and various other toxic waste products, which will be discussed more fully in Chapter 38.

The Mechanism for Excreting Large Quantities of Urea but Minimal Quantities of Water. Were it not for the ability of the tubular system to concentrate urea the kidneys could excrete the necessary amounts of urea only by passing large amounts of water into the urine at the same time. Instead, urea is normally concentrated more than 50-fold and, when a concentrated urine is being formed, sometimes as much as several hundred fold. Figure 35–6 illustrates this progressive increase in urea concentration in the tubular fluid as it passes through the tubular system. The numbers given in the diagram are those recorded during moderate antidiuresis, which means formation of small but highly concentrated amounts of urine. Note that the concentration of urea in the proximal tubule fluid is only 4.5 mOsm/liter, but this increases about 100 times (450 mOsm/liter) by the time the urea is excreted in the urine.

Recirculation of Urea from the Collecting Duct to the Loop of Henle as a Special Means for Concentrating the Urea. Note in Figure 35–6 the high concentration of urea in the lower collecting duct and also the rapid reabsorption of urea from this duct into the medullary interstitium, which increases the urea concentration there to high levels, to as high as 400 to 500 mOsm/liter. A large share of this urea is then reabsorbed into the thin limb of the loop of Henle so that it passes upward through the distal tubule, through the cortical collecting tubule, and back down through the collecting duct again. In this way, urea can recirculate through these terminal portions of the tubular system several times before it is excreted. And each time around this circuit contributes to the high concentrations shown in the figure so that very little water is excreted along with the urea.

Thus, this urea recirculation mechanism through the loop of Henle, the distal tubule, and the collecting tubule and duct is a way to concentrate urea in the medullary interstitium and in

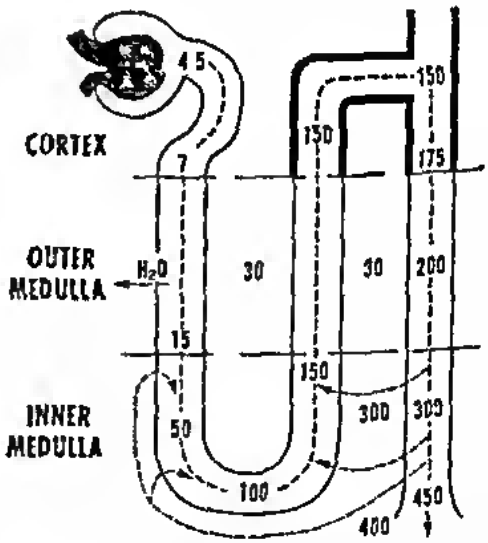


Figure 35–6. Recirculation of the urea absorbed from the collecting duct, this passing into the loop of Henle, through the distal tubule, and finally back to the collecting duct. (Numerical values are milliosmolalities of urea during antidiuresis caused by the presence of large amounts of antidiuretic hormone.)

the urine at the same time. Since urea is the most abundant of the waste products that must be excreted by the kidneys, this mechanism for concentrating urea before it is excreted is essential to the economy of the body fluid when a person must live on short rations of water.

SODIUM EXCRETION

The role of the kidney in sodium excretion is quite different from its role in urea excretion. The kidney attempts to eliminate as much urea as possible from the blood. To do this, large amounts of glomerular filtrate must be formed simply to supply the necessary tubular load of urea. By contrast, for sodium excretion the great amount of glomerular filtrate formed each day creates a serious problem. The reason for this is that the glomerular filtrate normally contains about 26,000 mEq of sodium each day, and yet the average intake of sodium per day is only 150 mEq. Therefore, the kidneys can be allowed to excrete only about 150 of the total 26,000 mEq. Consequently, the principal role of the tubular system in sodium excretion is to reabsorb sodium, not to excrete it. Furthermore, it must reabsorb just exactly the right amount of sodium, about 99.3 per cent of the total, and must adjust the final amount of sodium that issues into the urine hour by hour or day by day so that it will balance the daily intake of sodium. Thus, the two jobs that the tubular system must perform for sodium excretion are (1) to reabsorb nearly all of it, and (2) to adjust very carefully the remaining amount that is excreted for maintaining appropriate amounts of sodium in the body fluids.

Reabsorption of Most of the Tubular Sodium in the Proximal Tubules and Loops of Henle. By the time the tubular fluid reaches the distal tubules, all but about 8 per cent of the sodium has already been reabsorbed. About 65 per cent of this is reabsorbed in the proximal tubules because of the active transport of sodium by the proximal tubular epithelial cells. Also, as explained in the previous chapter, when the sodium is reabsorbed the positive charge of the sodium ions causes passive diffusion of negative ions, especially chloride ions, through the epithelium as well. Then the cumulative reabsorption of ions creates an osmotic pressure difference that moves water through the membrane, too. In fact, the epithelium is so permeable to water that almost the same identical proportion of water and sodium ions is reabsorbed.

In the thick portion of the ascending limb of the loop of Henle approximately another 27 per cent of the sodium is reabsorbed, leaving only 8 per cent to enter the distal tubules. However, reabsorption of sodium here is quite different from that in the proximal tubules because this thick segment of the loop is almost impermeable to water, as noted earlier in this chapter. Therefore, the concentration of sodium ions falls quite low in the ascending limb tubular fluid before it enters the distal tubules, down to as little as one quarter of its concentration in the plasma—to a concentration of 30 to 40 mEq per liter in contrast to the plasma level of 140 mEq per liter.

Thus, the proximal tubules and the loops of Henle are responsible for returning most of the sodium that enters the tubular system in the glomerular filtrate back to the plasma, thereby conserving the sodium for the body.

Variable Reabsorption of Sodium in the Late Distal Tubules and Cortical Collecting Tubules. Sodium reabsorption in the late distal tubules, the cortical collecting tubules, and the collecting ducts is highly variable. Its rate of reabsorption is controlled mainly by the concentration in the blood of *aldosterone*, a hormone secreted by the adrenal cortex. In the presence of large amounts of aldosterone, almost the last vestiges of tubular sodium are reabsorbed from these portions of the tubular system so that essentially none of the sodium enters the urine. On the other hand, in the absence of aldosterone, a large proportion of the sodium that enters the distal tubules is not reabsorbed and does then pass into the urine. Thus, the sodium excretion may be as little as 0.1 gram per day or as great as 30 grams per day, depending on the amount of aldosterone that is secreted.

Mechanism of Sodium Transport Through the Distal Tubule Collecting Tubule and Collecting Duct Epithelium. The mechanism for sodium transport through the tubular epithelium is illustrated in Figure 35–7. The movement is caused primarily by active transport of sodium from the interior of the tubular epithelial cell into the interstitial fluid of the lateral intercellular spaces and extensive channels at the base of the cell. The transport out of the cell creates a very low concentration of sodium inside the cell and also creates a negative electrical potential of approximately −70 millivolts within the cell. The low intracellular sodium concentration and the

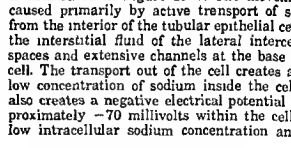


Figure 35–7. Mechanisms of Na^+ and K^+ transport through the distal tubular epithelium.

very negative electrical potential in turn create a high electrochemical gradient that causes sodium ions to diffuse from the tubular lumen into the cell, thus replacing the sodium as it is transported into the interstitial fluid.

The active transport of sodium out of the epithelial cell is coupled at least partially with active transport of potassium into the cell—potassium being exchanged for sodium, as we shall discuss further in relation to potassium excretion. For this reason, the excretion of sodium and potassium are to some extent coupled with each other, but in a reciprocal way.

Mechanism by Which Aldosterone Enhances Sodium Transport. Upon entering a tubular epithelial cell, aldosterone combines with a *receptor protein*; this combination diffuses within minutes into the nucleus where it activates the DNA molecules to form one or more types of messenger RNA The RNA is then believed to cause formation of carrier proteins or protein enzymes that are necessary for the sodium transport process. Various theories have suggested (a) that a specific protein increases the permeability of the luminal border of the cell to sodium, (b) that an increased quantity of sodium-potassium-ATPase (which is the sodium pumping protein) develops in the basolateral membrane of the epithelial cell, or (c) that a protein enzyme or enzymes increase the availability of ATP to the ATPase so that it can function more actively. Unfortunately, the precise mechanism is yet unknown.

Ordinarily, aldosterone has no effect on sodium transport for the first 45 minutes after it is administered, after this time the specific proteins important for transport begin to appear in the epithelial cells, followed by progressive increase in transport during the ensuing few hours

POTASSIUM EXCRETION

The amount of potassium entering the glomerular filtrate each day is about 800 mEq, whereas the daily intake of potassium is only about 100 mEq Therefore, to maintain normal body potassium balance, only one eighth of the total daily tubular load of potassium can be excreted. Furthermore, as is true for sodium excretion, the rate of potassium excretion must be carefully controlled so that it will exactly match the daily potassium intake

Absorption of Large Amounts of Potassium from the Proximal Tubules and Loops of Henle. Large amounts of potassium are reabsorbed in the same manner that sodium is reabsorbed in large quantities by the proximal tubules and the loops of Henle. Thus, active transport of potassium by the proximal tubular epithelial cells reabsorbs about 65 per cent of all the filtered ium Then, approximately another 27 per cent is reabsorbed by active transport in the thick portion of the ascending limb of the loop of Henle leaving about 8 per cent of the original filtered potassium to enter the distal tubules. This represents only about 65 mEq of potassium per day which is actually less than the average daily intake of potassium for most people.

Additional Reabsorption of Potassium by the Distal Tubules and Cortical Collecting Tubules. In addition to the absorption of potassium by the proximal tubules and loops of Henle, a small amount of active reabsorption of potassium occurs in the distal tubules and cortical collecting tubules, as illustrated by the solid arrow passing inward through the luminal membrane of the epithelial cell in Figure 35–7. Though this reabsorption is slight, nevertheless it is still sufficient to remove almost all the remaining tubular potassium. Therefore, if this were the only mechanism for potassium transport in the distal segments of the tubular system, essentially no potassium would be excreted in the urine, which would eventually be lethal because of the toxic effects of potassium accumulating in the body. Fortunately, though, the distal segments of the tubular system can also *secrete* potassium from the blood into the tubules, and this turns out to be the principal means by which the tubular system controls the rate of potassium loss in the urine. Therefore, let us discuss this secretion specifically.

Active Secretion of Potassium in the Late Distal Tubules and Collecting Tubules and Ducts. Normally, considerable amounts of potassium are secreted into the distal tubules, and some also into the collecting tubules and ducts. Figure 35–7 illustrates this secretion process. It shows that as sodium is transported from the cytoplasm of the epithelial cell into the peritubular fluid, potassium is simultaneously transported in the opposite direction to the interior of the cell. It is believed that this inward transport of potassium is not rigidly coupled with the outward transport of sodium. Instead, either all or most of the potassium movement into the cell is probably caused by the very negative electrical potential created inside the cell when sodium is pumped out into the interstitial fluid—this negativity attracting the positively charged potassium ions into the cell. Once the potassium has entered the epithelial cell, it then diffuses passively from the cell into the tubular lumen.

In the absence of tubular secretion of potassium, death will eventually ensue from potassium toxicity. Indeed, cardiac arrhythmias usually appear when the plasma potassium concentration rises from the normal value of 4.5 mEq per liter to a level of 8 mEq per liter. A still higher potassium concentration than this can end in cardiac arrest or fibrillation

Control of the Rate of Potassium Secretion. The rate of tubular secretion of potassium is con-

trolled in proportion to the need to eliminate potassium from the extracellular fluid. To achieve this, three major factors determine the rate of tubular secretion of potassium:

(1) Control of potassium secretion by aldosterone. Earlier, we noted that aldosterone has a potent effect in controlling sodium transport through the basolateral membrane of the epithelial cells in the terminal portions of the tubular system, beginning in the late distal tubules. And it was also pointed out that sodium being transported out of the cell into the interstitium causes potassium transport from the interstitium into the cell and then diffusion of the potassium through the luminal membrane on into the tubule. Therefore, aldosterone controls potassium secretion at the same time that it controls sodium reabsorption Indeed, for peculiar reasons that will be presented in the following chapter, when we discuss the control of the concentrations of sodium and potassium ions in the extracellular fluids, we shall see that aldosterone is actually much more important for controlling extracellular fluid potassium concentration than for controlling sodium concentration

(2) Effect of plasma potassium concentration on rate of potassium secretion When the plasma potassium concentration rises too high, the rate of potassium transport into the tubular lumen increases markedly. The probable reason for this is that the increased potassium concentration gradient from the interstitial fluid to the tubular lumen now allows extra potassium ions to be attracted into the tubular cells by the very negative electrical potential inside the cells that occurs when sodium ione ere pumped out of the cells into the interstitium. Once inside the cells, the potessium lons diffuse Into the tubular lumen. Regardless of the cause, this marked enhancement of potassium secretion is another important factor in the control of extracellular fluid potassium concentration.

(3) Effect of distol tubulor sodium load on potassium secretion. When the quantity of sodium entering the late distal tubules increases, this normally increases the rate of sodium reabsorption by the late distal tubules and collecting tubules. And, because potassium secretion is inversely coupled to sodium reabsorption, this has the indirect effect of increasing the rate of potassium secretion. Therefore, the available quantities of sodium entering the distal tubules are an additional factor that enters into the control of potassium secretion as well as into the control of extracellular fluid potassium ion concentration.

FLUID VOLUME EXCRETION

Up to this point we have considered the intrarenal mechanisms that determine the *concentrations* of various substances in the urine—water, urea, sodium, and potassium. Now it is important to consider the different factors that determine the rate of fluid *volume* excretion.

Importence of Conserving Body Fluid, and the Concept of "Glomerulotubular Balance." The kidney's problem in fluid volume excretion is much the same as for sodium excretion, that is, not to lose too much volume from tbe body. To achieve this, over 99 per cent of the glomeruler filtrete volume must be reabsorbed by the tubules. Also, when the glomerular filtration rate increases, the tubules, especially the proximal tubules, automatically adjust their rates of volume reabsorption almost to keep step with the increased rate of filtration. This phenomenon is called "glomerulotubular balance."

To say that there is exact *glomerulotubular balance* means that when the rate of glomerular filtration increases, so also does the rate of tubular reabsorption increase in exact proportion to the increase in filtration. That is, if the entry of glomerular filtrate into the proximal tubules increeses from 60 to 120 ml per minute, a 100 per cent increese, exact proximal tubuler glomerulotubular balence would be achieved if the rete of reabsorption in the proximal tubules should elso increase 100 per cent, from approximately 40 to 80 ml per mlnute, thus maintaining approximetely the same percentage of reabsorption regardless of the rate of fluid entry into the tubular system.

The reason that the glomerulotubular belence principle holds true (or at least neerly holds true) is that increesed flow of filtrate into the tubuler system markedly ectivates most of the tubular reabsorptive mechanisms so thet they almost keep up with the increesed volume of fluid to be processed.

Glomerulotubular balance is especially good in the proximal tubules and loop of Henle, but somewhat lesa effective in the more distal segments of the tubular segment. And this lack of 100 per cent glomerulotubular balance turns out to be an especially important factor in determining urine volume output, as we shell explain.

Glomerulotubular "Imbalance" and Its Importance to Fluid Volume Regulation. Though the concept of glomerulotubular balance is important to explain the way in which glomerular filtration rate and rate of tubular reabsorption automatically adjust to one another, very precise measurements show that 100 per cent glomerulotubular balance for the entire tubular system almost never occurs. For instance, the following table gives approximate values for glomerular filtration rates, total rates of tubular fluid reabsorption, and rates of urine output for the average human adult, as extrapolated mainly from data in dogs:

Glomerular Filtration Rate ml	Rate of Tubular Reabsorption ml	Rate of Urine Output ml
50	49.8	0 2
75	74.7	0.3
100	99.5	0.5
125	124.0	1.0
150	145.0	5.0
175	163.0	12.0

If we examine these figures critically, we see that glomerular filtration rate and rate of tubular reabsorption actually do appear to parallel each other very closely. On the other hand, the slight degree of imbalance that does occur causes far greater change, proportionately, in urine output than in either glomerular filtration rate or tubular reabsorption rate. For instance, let us study the increase in glomerular filtration rate from 100 to 150 ml/min. The rate of reabsorption increases from 99.5 to 145 ml/min, representing only slight glomerulotubular imbalance. Nevertheless, this 50 per cent increase in glomerular filtration rate causes a 1000 per cent increase in rate of urine output. Thus, even a very slight degree of glomerulotubular imbalance can lead to a tremendous increase in urine output when the glomerular filtration rate is increased. Also, very slight changes in rate of reabsorption of tubular fluid can cause equally as great alterations in urine output.

Therefore, the various factors that can alter either glomerular filtration rate or rate of tubular reabsorption are also the factors that play significant roles in determining the rate of fluid volume excretion. The five most important of these are: (1) the tubular osmolar clearance, (2) the plasma colloid osmotic pressure, (3) the degree of sympathetic stimulation of the kidneys, (4) the arterial pressure, and (5) the effect of antidiuretic hormone on tubular reabsorption of water.

1. Effect of Tubular Osmolar Clearance on Rate of Fluid Volume Excretion

Under most normal conditions, approximately proportional quantities of solutes and water are reabsorbed from the tubules. That is, whenever an osmolar substance is reabsorbed, this increases the osmotic gradient across the tubular wall and causes almost proportionate osmosis of water from the tubules into the interstitial fluid. Conversely, the greater the quantity of osmolar substances that *fails* to be reabsorbed by the tubules, the greater the quantity of water that also fails to be reabsorbed. To state this another way, when the osmolar clearance is great, the volume of urine usually increases by approximately the same percentage.

Osmotic Diuresis. The effect of increased osmolar clearance in causing a proportionate amount of water to be lost in the urine is called *osmotic diuresis*. A particularly interesting type of osmotic diuresis occurs in diabetes mellitus in which the proximal tubules fail to reabsorb all the glucose as normally occurs. Instead, the nonreabsorbed glucose passes the entire distance through the tubules and carries with it a large portion of the tubular water. Therefore, in diabetes mellitus (the word "diabetes" means diuresis) the urine output occasionally increases to as high as 4 to 5 liters per day.

Osmotic diuresis also occurs when substances that cannot be reabsorbed by the tubules are filtered in excessive quantities from the plasma into the glomerular filtrate. For instance, sucrose, mannitol, and urea, when in the circulating plasma in large quantities, are all filtered into the glomerular filtrate and cause large tubular loads of osmotic substances that either are not reabsorbed at all or are reabsorbed very poorly. Therefore, they tend to cause extreme osmotic diuresis.

2. Effect of Plasma Colloid Osmotic Pressure on Rate of Fluid Volume Excretion

Another factor that greatly affects the rate of volume excretion is the plasma colloid osmotic pressure. A sudden increase in plasma colloid osmotic pressure instantaneously decreases the rate of fluid volume excretion. The cause of this is twofold: (1) an increase in plasma colloid osmotic pressure *decreases* glomerular filtration rate, and (2) an increase in plasma colloid osmotic pressure *increases* tubular reabsorption. Both of these effects add together to decrease greatly the urine volume excretion.

Decreased Glomerular Filtration Rate. An increase in plasma colloid osmotic pressure of 7 mm Hg decreases glomerular filtration rate as much as 5 to 10 per cent. The cause of this is simply the decrease in glomerular filtration pressure.

Increased Tubular Reabsorption. The effect of increased plasma colloid osmotic pressure to increase tubular reabsorption is less well understood, but the mechanism is believed to be that illustrated in Figure 35–8. This figure shows transport of sodium and water from the proximal tubule into the spaces surrounding the epithelial cells. Once the sodium and water (and the other substances that move along with these two) have entered the intercellular spaces and the labyrinthine channels at the bases of the epithelial cells, the fluid can then go one of two ways: either it can leak backward into the tubule through the junctions between the epithelial cells, or it can move forward through the basement membrane into the peritubular capillary. (Remember that the so-called "tight" junctions between the epithelial cells of the proximal tubules are loose enough to allow considerable back-leakage of sodium ions and water.)

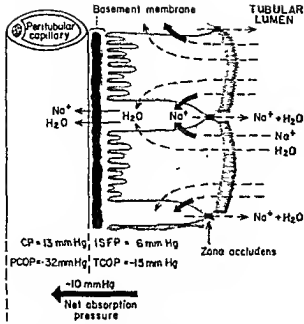


Figure 35–8. Absorption of fluid at the peritubular capillary membrane, and the effect of this on the absorption of water and sodium through the tubular epithelium.

One of the major factors that determines whether or not the sodium and water will leak backward into the tubular lumen or will move forward into the peritubular capillary is the balance of hydrostatic and colloid osmotic pressures in the capillary membrane At the bottom of Figure 35–8 the normal balance of pressures is depicted. Any one of the four factors affecting the balance at the capillary membrane changes, this obviously will also change the net absorption pressure. Thus, an increase in the plasma colloid osmotic pressure will increase the rate at which the capillary pulls fluid through the basement membrane, which also decreases the rate at which fluid leaks backward into the tubular lumen. To a very great extent, therefore, the rate of fluid absorption from the tubular lumen into the peritubular capillaries is determined by the plasma colloid osmotic pressure.

3. Effect of Sympathetic Stimulation on Rate of Fluid Volume Excretion

Sympathetic stimulation has an especially powerful effect on constriction of the afferent arterioles. It greatly decreases the glomerular pressure and simultaneously decreases glomerular filtration rate. And, at the same time, decreased blood flow into the peritubular capillaries decreases the capillary pressure, thus increasing tubular reabsorption.

Conversely, a decrease of sympathetic stimulation to below normal causes a mild degree of afferent arteriolar dilatation, which increases the glomerular filtration rate slightly and also increases peritubular capillary pressure which decreases reabsorption. Consequently, decreased sympathetic stimulation leads to increased urine volume excretion.

4. Effect of Arterial Pressure on Rate of Fluid Volume Excretion

If all other factors remain constant but the renal arterial pressure is changed, the rate of urine output changes markedly. This effect is illustrated in Figure 35–9, showing that when the arterial pressure rises from 100 to 200 mm Hg the increase in urine output is approximately sevenfold. Conversely, when the arterial pressure falls from 100 to 60 mm Hg, the urine output falls either to zero or near zero. We have already pointed out in Chapter 22 that this pressure effect on urine output plays an extremely important role in the feedback regulation of arterial pressure. It also plays an extremely important role in the feedback regulation of body fluid volume, as we shall discuss in the following chapter. Now, however, let us discuss the mechanism of the increased urine output caused by the increase in arterial pressure. This results from two separate effects: (1) the increase in arterial pressure increases glomerular pressure, which in turn increases glomerular filtration rate, thus leading to increased urine output; (2) the increase in arterial pressure also increases the peritubular capillary pressure, thereby decreasing tubular reabsorption. The combination of these two effects causes considerable glomerulotubular imbalance and therefore also causes marked increase in urine output, as illustrated in Figure 35–9.

5. Effect of Antidiuretic Hormone on Rate of Fluid Volume Excretion

When excess antidiuretic hormone is secreted by the posterior pituitary gland, the effect is to decrease the urinary volume output acutely. The

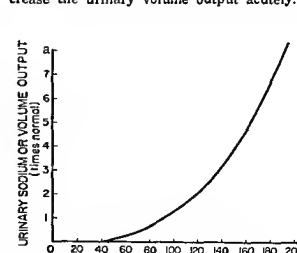


Figure 35–9. Effect of arterial pressure change on urinary output.

reason for this is that antidiuretic hormone causes increased water reabsorption from the collecting tubules and ducts and in some species of animals from the late distal tubules as well. Therefore, less urinary volume is excreted; on the other hand, the urine that is excreted is highly concentrated, as was explained earlier in the chapter.

When excess antidiuretic hormone is secreted for long periods of time, the acute effect of decreasing urinary output is not sustained. The reason is that other factors, such as the arterial pressure, colloid osmotic pressure, and concentrations of the osmolar substances in the glomerular filtrate all change in the direction that leads eventually to a urinary volume output equal to the daily intake of fluid minus the losses of fluid in other ways from the body. In other words, balance between body fluid intake and body fluid output is reestablished. Therefore, contrary to what is often taught, long-term secretion of excess antidiuretic hormone plays only a small role in chronic regulation of body fluid volume. As we shall see in discussion of the condition called "inappropriate ADH syndrome" in the following chapter, continued long-term secretion of very large quantities of antidiuretic hormone causes severe decreases in extracellular sodium, potassium, and chloride concentrations even though the body fluid volume is only slightly altered.

SUMMARY OF THE CONTROL OF FLUID VOLUME EXCRETION

From the preceding few sections, it is clear that many different factors contribute to the regulation of urine volume excretion. Some of the factors cause extreme and acute changes in urinary output—for instance, the acute decrease in urine volume output caused by antidiuretic hormone or the acute increase in urine volume output caused by increased tubular osmotic loads. However, over a longer period of time, other longer-acting factors, such as changes in arterial pressure and colloid osmotic pressure, tend to override the extreme, acute effects of these first factors

REFERENCES

Anagnostopoulos, T., *et al* Transport of chlorine in the proximal tubule Its effects on water-electrolyte absorption *J Physiol (Paris)*, *79* 132, 1984

Arendshorst, W J., and Gottschalk, C W. Glomerular ultrafiltration dynamics Historical perspective *Am J Physiol.*, *248* F163, 1985

Aukland, K · Renal blood flow *Int Rev Physiol*, *11* 23, 1976

Aukland, K Methods for measuring renal blood flow Total flow r regional distributions *Annu Rev Physiol*, *42* 543, 1980

Blantz, R C. Segmental renal vascular resistance Single nephr *Annu Rev Physiol*, *42* 573 1980

Braun, E J, and Dantzler, W H: Endocrine regulation of avian re function *J Exp Zool*, *232* 715, 1984

Brenner, B M. Adaptation of glomerular forces and flows to re injury *Yale J Biol Med*, *51* (3) 301, 1978

Brenner, B M, *et al*. Determinations of glomerular filtration r *Annu Rev Physiol*, *39* 9, 1976

Brenner, B M, *et al*: Transport of molecules across renal glomerul capillaries *Physiol Rev*, *56* 502, 1976

Brucker, N S The Kidney Diagnosis and Management. New Yo John Wiley & Sons, 1984

Catto, G R Clinical Aspects of Renal Physiology Philadelphia W B Saunders Company, 1991

Edmonds, C J Salts and water *Biomembranes*, *4B* 111, 1974

Giebisch, G and Stanton, B Potassium transport in the nephr *Annu Rev Physiol*, *41* 241, 1979

Goldberg M *et al* Renal handling of calcium and phosphate *Int R Physiol* *11* 211, 1976.

Hall, J E, *et al* A single injection method for measuring glomerul filtration rate *Am J Physiol*, *232* F72, 1977

Hall, J E *et al* Dissociation of renal blood flow and filtration r autoregulation by renin depletion. *Am J Physiol*, *232* F215 1977

Herholzer K, and Lange, S The effects of adrenal steroids on re function *In* MTP International Review of Science Physiology Vol Baltimore, University Park Press, 1974, p 273

Ives, H E and Rector, F C, Jr Proton transport and cell function *Clin Invest*, *73* 285 1984

Jamison, R L., and Hall, D A Collecting duct function and sodiu balance *Annu Rev Physiol*, *33* 241, 1982

Kanwar, Y S Biophysiology of glomerular filtration and proteinur *Lab Invest*, *51* 7, 1984

Kenny A. J and Maroux, S Topology of microvillar membra hydrolases of kidney and intestine *Physiol Rev.*, *62* 91, 1982

Knox F G and Diaz Suxo, J A. The hormonal control of sodiu excretion *Int Rev Physiol*, *16* 173, 1977

Kramer, H J, and Krück, F (eds.) Natriuretic Hormone New Yo Springer Verlag, 1978

Lohmeier, T E, *et al* Effects of endogenous angiotensin II on re sodium excretion and renal hemodynamics *Am J Physiol*, *233* F3 1977

Moses, A M and Share, L, (eds) Neurohypophysis. New York Karger 1977

Navar, L G, *et al* Effect of alterations in plasma osmolality on re blood flow autoregulation *Am J Physiol.*, *211* 1337, 1966

O Connor, W J Normal Renal Function. The Excretion of Water a Electrolytes Derived From Food and Drink New York Oxford Un versity Press, 1982

Renkin, E M, and Robinson, R R Glomerular filtration. *N Engl Med*, *290* 785, 1974

Schafer, J A and Andreoli, T E Rheogenic and passive Na+ absor tion by the proximal nephron *Annu Rev Physiol*, *41* 211, 1979

Schrier, R W Effects of adrenergic nervous system and catecholamin on systemic and renal hemodynamics, sodium and water excretio and renin excretion *Kidney Int.*, *6* 291, 1974

Smith, W L., and Garcia-Perez, A Immunodissection Use of mon clonal antibodies to isolate specific types of renal cells. *Am J Physio* *248* F1, 1985

Stephenson, J L. Countercurrent transport in the kidney *Annu. R Biophys Bioeng*, *7* 315, 1978

Weitzman, R, and Kleeman, C R.: Water metabolism and the neur hypophysial hormones. *In* Bondy, P K., and Rosenberg L. E (eds Metabolic Control and Disease. 8th Ed Philadelphia, W B Saunde Co 1980, p 1241

Wright, F S Intrarenal regulation of glomerular filtration rate *Engl J Med*, *291* 135, 1974

Wright, F S, and Briggs, J P: Feedback control of glomerular blo flow, pressure and filtration rate *Physiol Rev.*, *59* 958, 1979

36

Regulation of Blood Volume, Extracellular Fluid Volume, and Extracellular Fluid Composition by the Kidneys and by the Thirst Mechanism

The kidneys, more than any other organ, play letermining roles in regulating important charicteristics of the body fluids, including (1) blood volume, (2) extracellular fluid volume, (3) osmoality of the body fluids—that is, the ratio of water o dissolved substances, (4) specific concentrations of the different ions, and (5) degree of acidity the body fluids. The first four of these will be liscussed in the present chapter, and the last—regulation of hydrogen ion concentration—will be liscussed in the following chapter. In controlling some of these characteristics the thirst mechanism also plays a key role; this will also be discussed.

CONTROL OF BLOOD VOLUME

Constancy of the Blood Volume. The extreme degree of precision with which the blood volume is controlled is illustrated in Figure 36–1. This shows the effect of changing the daily fluid intake, including both water end dissolved electrolytes, from very low values to very high values. It indicates that almost no change in blood volume occurs despite tremendous changes in intake, except when the intake becomes so low that it is not sufficient to make up for fluid losses caused by evaporation or other inescapable losses.

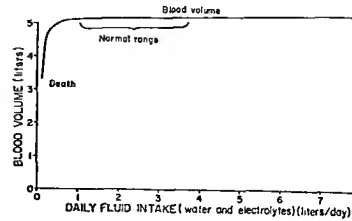


Figure 36–1. Effect on blood volume of marked changes in daily fluid intake. Note the precision of blood volume control in the normal range.

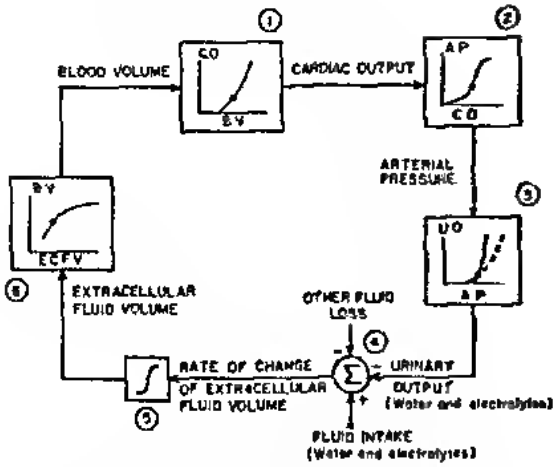


Figure 36–2. Basic feedback mechanism for control of blood volume and extracellular fluid volume (the points on each of the curves represent normal values)

BASIC MECHANISM FOR BLOOD VOLUME CONTROL

The basic mechanism for blood volume control is illustrated in Figure 36–2. This is essentially the same as the basic mechanism for arterial pressure control that was presented in Chapter 22. It was pointed out in that chapter that extracellular fluid volume, blood volume, cardiac output, arterial pressure, and urine output are all controlled at the same time as separate parts of a common basic feedback mechanism. The six major steps in this mechanism as they relate to blood volume control are:

Block 1 shows that an increase in *blood volume increases cardiac output.*

Block 2 shows that an increase in cardiac output increases *arterial pressure.*

Block 3 shows that an increase in *arterial pressure increases urinary output,* the solid curve in this block indicates that the effect of pressure on the output is especially potent when the pressure remains elevated for days or weeks at a time.

Block 4 gives a *summation of the fluid intake minus the fluid losses* from the body, the losses including both urinary output and other fluid loss. The output of this block is the *rate of change of extracellular fluid volume.* If the intake is greater than the output, the rate of change will be positive; if the output is greater than the intake, the rate of change will be negative.

Block 5 *integrates the rate of change of extracellular fluid volume*—that is, it shows the accumulation of greater or less fluid volume with time, depending on whether the rate of change is positive or negative. The output of Block 5 is the actual extracellular fluid volume.

Block 6 gives the *relationship between extracellular fluid volume and blood volume,* showing that, in general, *as the extracellular fluid volume increases, the blood volume also increases.*

Summary of the Basic Blood Volume Control Mechanisms. To summarize the principles illustrated in Figure 36–2, we can trace what happens when the blood volume becomes abnormal. When the blood volume becomes too great the cardiac output also becomes too great and therefore increases the arterial pressure. This, in turn, has a profound effect on the kidneys, causing loss of fluid from the body and returning the blood volume back to normal. Conversely, if the blood volume falls below normal, the cardiac output and arterial pressure decrease, the kidneys retain fluid and progressive accumulation of the fluid intake builds the blood volume eventually back to normal. Usually, parallel processes will also occur to reconstitute red cell mass, plasma proteins, and so forth if these have become abnormal at the same time.

However, if the red cell volume remains abnormal, the plasma volume will simply make up the difference, the volume becoming essentially normal despite the low red cell mass.

Reason for the Precision of the Blood Volume-Regulating Mechanism. Studying the diagram of Figure 36–2 very carefully, one can see why the blood volume remains almost exactly constant despite extrema changes in daily fluid intake. The reason for this is that the slopes of the curves in Blocks 1, 2, and 3 are all very steep, meaning that a slight change in blood volume causes a marked change in cardiac output, a slight change in cardiac output causes a marked change in arterial pressure, and a slight change in arterial pressure causes a marked change in urinary output. These factors all multiply together to give an extremely high gain for the feedback control of blood volume.

ROLE OF THE VOLUME RECEPTORS IN BLOOD VOLUME CONTROL

It was pointed out in Chapter 21 that "volume receptor" reflexes help control blood volume. These modify the function of the basic mechanism for blood volume control in a very specific way, as follows:

The basic mechanism of Figure 36–2 is slow to come to equilibrium. That is, when the blood volume increases, the effects on cardiac output, arterial pressure, and urinary output are slow to build up, often requiring many hours to develop full effect; even after they do develop full effect, additional hours are still required for the excess fluid to be eliminated from the body. However, the volume receptor reflexes can greatly accelerate this process.

The volume receptors are mainly stretch receptors located in the walls of the right and left atria. When the blood volume becomes excessive, a large share of the extra volume accumulates in the central veins of the thorax and causes increased pressure in the two atria. The resultant stretch of the atrial walls transmits nerve signals into the brain, and these in turn elicit two important renal responses that accelerate the return of blood volume to normal. These responses are:

1. The sympathetic nervous signals to the kidneys are

inhibited, which causes a moderate increase in the rate of urinary output.

2. The signals from the atrial stretch receptors also inhibit the secretion of antidiuretic hormone by the posterior pituitary gland, and lack of this hormone allows the kidneys to excrete increased quantities of water.

In most instances, these volume receptor reflex effects can cause the blood volume to return almost all the way to normal within an hour or so, but the final determination of the precise level to which the blood volume will be adjusted is still a function of the basic volume control mechanism illustrated in Figure 36–2. The reason for this is that over a period of one to three days the volume receptors adapt completely so that they no longer transmit any corrective signals Therefore, they are of value only to help readjust the volume during the first few hours or few days after an abnormality occurs, but not for long-term monitoring of volume or for precise adjustments of the long-term level of blood volume.

Role of the Baroreceptors and Other Stretch Receptors in the Volume Receptor Response. Ordinarily, when the blood volume increases, the systemic arterial pressure and the pulmonary arterial pressure also increase at least to some extent Therefore, the baroreceptors located in the walls of the carotid, aortic, and pulmonary arteries are excited These, too, cause essentially the same reflex effects as the atrial volume receptors. Therefore, the baroreceptors add still more to the volume receptor reflex effect

Possible Role of "Atrial Natriuretic Factor" in Volume Control. Large numbers of cells in the walls of the two atria, but especially in the right atrium, contain a hormone-like substance called *atrial natriuretic factor*. When this factor is infused into the circulating blood, it causes as much as a three to ten times increase in renal excretion of sodium And, concomitant with the sodium loss, both the extracellular fluid volume and blood volume decrease slightly Therefore, it has been suggested that atrial natriuretic factor might play a role in the regulation of blood volume. The theory is that excess blood volume stretches the atrial walls, and this promotes the release of atrial natriuretic factor that in turn causes sodium excretion and reduction of the blood volume. Unfortunately, the quantitative aspects of this mechanism are yet unclear, but it could operate in parallel with the volume receptor mechanism to help regulate blood volume.

OTHER FACTORS THAT HELP CONTROL BLOOD VOLUME

Capacity of the Circulation. Any change in the basic capacity of the circulatory system will automatically change the level to which the blood volume is regulated. For instance, when a person develops severe varicose veins, the total capacity of the circulatory system may increase as much as 1 liter. And the volume-controlling mechanisms will automatically control the blood volume to this higher level.

Other factors that sometimes change the capacity of the system include vasoconstrictor factors, vasodilator factors, aneurysms, and so forth For instance, when a person is under the chronic influence of strong sympathetic vasoconstrictor stimulation or of vasoconstrictor hormones such as norepinephrine, the blood volume becomes regulated to a lower level. That is, the volume control system is geared to adjust to the volume of blood to fill the capacity of the system itself.

Long-Term Effects of Antidiuretic Hormone and Aldosterone on Blood Volume. The hormones antidiuretic hormone and aldosterone both play important roles in some aspects of fluid and electrolyte economy of the body. Therefore, both of them have been extolled as very important blood volume regulators. However, measurements show that these normally affect blood volume relatively little, except for the temporary effect of antidiuretic hormone in helping adjust the blood volume when the "volume receptors" are stimulated, as noted earlier. Therefore, let us discuss only briefly their effects on blood volume regulation.

Antidiuretic Hormone. When the posterior pituitary gland fails to secrete antidiuretic hormone, water absorption in the distal segments of the renal tubular system is greatly decreased, thus allowing urine output to increase as much as three to tenfold. This condition is called *diabetes insipidus* Despite this extreme increase in urine output, the blood volume normally does not decrease enough to be measurable. The reason is that the thirst mechanism simply causes the person to drink enough water to make up the difference. On the other hand, if the person is prevented from obtaining water to drink, then the blood volume is likely to decrease considerably, and this can cause circulatory shock However, this combination only rarely occurs.

Occasionally, antidiuretic hormone is secreted in tremendous amounts. This condition is called *syndrome of inappropriate ADH secretion*. However, here again, in the chronic state of this condition the blood volume increases almost imperceptibly—perhaps as much as 3 to 5 per cent. The reason for this minute increase is that the slight volume increase that does occur increases the arterial pressure enough to overcome the antidiuretic hormone's effect on the kidneys of causing water retention. Experiments have shown that only a few millimeters of mercury rise in pressure—actually too little to be measured without special care—causes enough pressure diuresis to overcome the antidiurstic effect of antidiuretic hormone

Aldosterone. Aldosterone causes excessive salt reabsorption from the late distal tubules, collecting tubules, and collecting ducts of the kidneys, and this in turn causes osmotic reabsorption of water. Therefore, the immediate effect is to decrease urine output greatly. The extracellular fluid volume and the blood volume both begin to increase However, before these can increase more than a few per cent, the basic feedback mechanism for blood volume control (see Figure 36–2) comes into play and overbalances the retention of fluid by the kidneys. It does this by raising the arterial pressure a few millimeters of Hg, increasing glomerular filtration rate a few per cent, and therefore increasing the urinary output back to the point at which it equals the intake of fluid.

Consequently, even in persons who have tremendous secretion of aldosterone (patients with *primary aldosteronism*), the extracellular fluid volume and blood volume rarely rise more than 5 to 10 per cent, at most. On the other hand, in persons who have no secretion of aldosterone (patients with *Addison's disease*) the late distal tubules, collecting tubules, and collecting ducts fail to reabsorb salt and water, and the kidneys lose tremendous quantities of fluid into the urine. If the enough salt and drinks a concomitant the blood volume will still be regulated

range. But if salt and water intakes are not sufficient, the patient can develop severe dehydration and hypovolemia with resultant circulatory shock

CONTROL OF EXTRACELLULAR FLUID VOLUME

It is already clear from the foregoing discussion of the basic mechanisms for blood volume control that extracellular fluid volume is controlled at the same time. That is, fluid first goes into the blood, but it rapidly becomes distributed between the interstitial spaces and the plasma Therefore, it is impossible to control blood volume to any given level without controlling the extracellular fluid volume at the same time. Yet, the relative volumes of distribution between the interstitial spaces and the blood can vary greatly, depending on the physical characteristics of the circulatory system and of the interstitial spaces. Under normal conditions, the fluid in the interstitial spaces is bound in a gel-like matrix of hyaluronic acid molecules, and there is essentially no free fluid. At other times, however, abnormal conditions can cause edema to occur. These abnormal conditions were discussed in detail in Chapter 31. The principal factors that can cause edema are (1) increased capillary pressure, (2) decreased plasma colloid osmotic pressure, (3) increased tissue colloid osmotic pressure, or (4) increased permeability of the capillaries Whenever any one of these conditions occurs, a very high proportion of the extracellular fluid becomes distributed to the interstitial spaces.

Normal Distribution of Fluid Volume Between the Interstitial Spaces and the Vascular System. Figure 36–3 illustrates the approximate normal relationship between extracellular fluid volume and blood volume When an extra amount of fluid accumulates in the extracellular fluid space, either as a result of intake of too much flui or because of decreased kidney output of flui about one sixth to one third of the extra flu normally stays in the blood and increases the bloo volume. The remainder of the fluid is distribute to the interstitial spaces. However, when the ex tracellular fluid volume rises considerably abo normal, there comes a point, as shown in th figure, beyond which very little of the addition fluid will remain in the blood—almost all of instead going into the interstitial spaces. Th occurs when the interstitial fluid pressure ris from its normal negative (subatmospheric) valu to a positive value, because only a few millimete of Hg can cause the tissues to swell so greatly th they may then hold as much as 10 to 30 liters extra fluid. That is, the interstitial spaces literal become an "overflow" reservoir for excess flui This obviously causes edema, as was explained i Chapter 31, but it also acts as an important ove flow release valve for the circulatory system, well-known phenomenon that is utilized daily b the clinician to allow administration of almo unlimited quantities of intravenous fluid and y not to force the heart into cardiac failure.

To summarize, extracellular fluid volume is co trolled simultaneously with the control of bloo volume, but the relative ratio of the extracellul fluid volume to blood volume depends upon th physical properties of the circulation and of th interstitial spaces, and especially on the dynami of fluid exchange through the capillary mem branes.

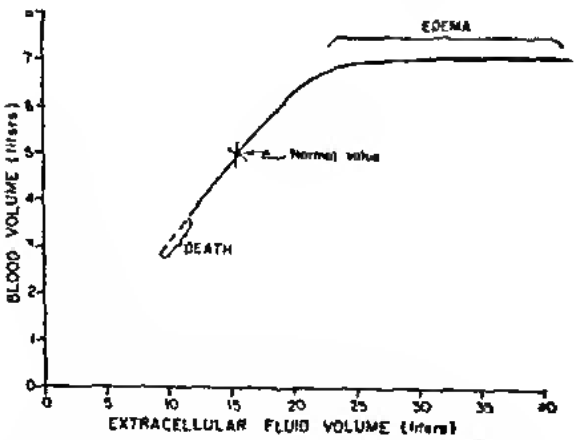


Figure 36–3. Relationship between extracellular fluid volume and blood volume, showing a nearly linear relationship in the normal range but indicating failure of the blood volume to continue rising when the extracellular fluid volume becomes

CONTROL OF EXTRACELLULAR FLUID SODIUM CONCENTRATION AND EXTRACELLULAR FLUID OSMOLALITY

The sodium ion concentration of the extracell lar fluid is normally about 142 mEq/liter; th rarely rises or falls more than 2 to 3 per cen above or below this level, illustrating the extrem effectiveness of the control systems for regulatin the extracellular fluid sodium concentration Fu thermore, the osmolality of the extracellular flui is normally very near 300 mOsm per liter, an this too rarely changes more than ± 3 per cen because the regulation of sodium ion concentratio and extracellular fluid osmolality are very tightl linked to each other, as will be explained in th following few sections.

Relationship of Sodium Concentration t Extracellular Fluid Osmolality. The osmolalit of the extracellular fluids is determined almo entirely by the extracellular fluid sodium conce tration. The reason for this is that sodium is b far the most abundant positive ion of the extrace lular fluid, representing over 90 per cent of the ions. Furthermore, various renal control mech

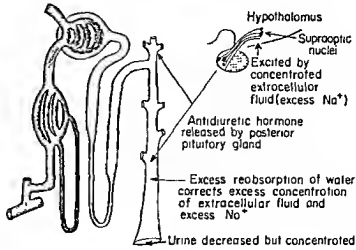


Figure 36–4. Control of extracellular fluid osmolality and sodium ion concentration by the osmosodium receptor-antidiuretic hormone feedback control system

isms, especially the fact that every time a positive ion is reabsorbed a negative ion also is reabsorbed, adjust the negative ion concentration of the body fluids to equal that of the positive ions. Therefore, in effect, the control of the positive ions controls the total ion concentration. Also, the glucose and urea, which are the most abundant of the nonionic osmolar solutes in the extracellular fluids, normally represent only 3 per cent of the total osmolality, and even then the urea exerts very little *effective* osmotic pressure because it penetrates cells too easily to cause significant osmotic results. Therefore, in effect, *the sodium ions of the extracellular fluid determine either directly or indirectly over 90 per cent of the osmotic pressure of the extracellular fluid.* Consequently, we can generally talk in terms of control of sodium concentration and control of osmolality at the same time.

Two separate control systems operate in close association to regulate extracellular sodium concentration and osmolality. These are (1) the osmosodium receptor–antidiuretic hormone system, and (2) the thirst mechanism

THE OSMOSODIUM RECEPTOR–ANTIDIURETIC HORMONE FEEDBACK CONTROL SYSTEM

Figure 36–4 illustrates the osmosodium receptor–antidiuretic hormone system for control of extracellular fluid sodium concentration and osmolality. It is a typical feedback control system that operates by the following steps:

(1) An increase in osmolality (mainly excess sodium and the negative ions that go with it) excites *osmoreceptors* located in the supraoptic nuclei of the hypothalamus.

(2) Excitation of the supraoptic nuclei causes the posterior pituitary gland to release *antidiuretic hormone.*

(3) The antidiuretic hormone increases the permeability of the lato distal tubules, the collecting tubules, and the collecting ducts, as explained in the previous chapter, and therefore *causes increased conservation of water by the kidneys.*

(4) The conservation of water but *loss of sodium and other osmolar substances in the urine* causes dilution of the sodium and other substances in the extracellular fluid, thus correcting the initial, excessively concentrated extracellular fluid.

Conversely, when the extracellular fluid becomes too dilute (hypo-osmotic), less antidiuretic hormone is formed, and excess water is lost in comparison with the extracellular fluid solutes, thus concentrating the body fluids back toward normal.

The Osmoreceptors (or Osmosodium Receptors). Located in the supraoptic nuclei of the anterior hypothalamus, shown in Figure 36–5, are specialized neuronal cells called *osmoreceptors.* These respond to changes in osmolality of the extracellular fluid. When the osmolality of the extracellular fluid becomes low, osmosis of water occurs into the osmoreceptors, causing them to swell. This decreases their rate of impulse discharge. Conversely, increased osmolality in the extracellular fluid pulls water out of the osmoreceptors, causing them to shrink and thereby to increase their rate of discharge.

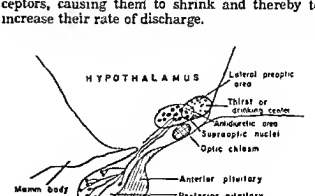


Figure 36–5. The supraopticopituitary antidiuretic system and its relationship to the thirst center in the hypothalamus.

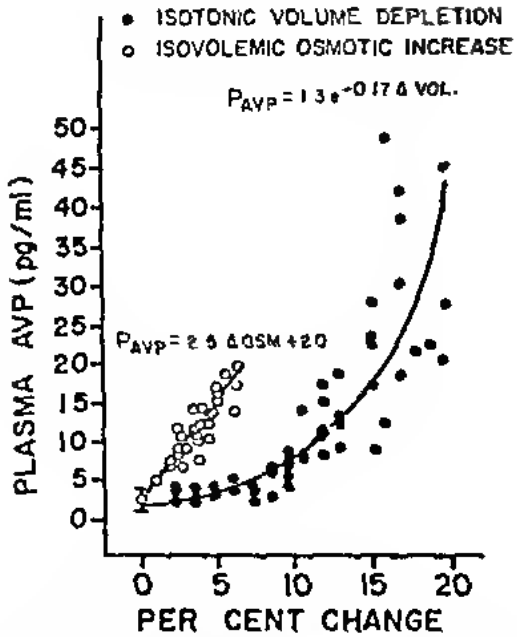


Figure 36–6. Effect of changes in plasma osmolality or blood volume on the level of plasma ADH (arginine vasopressin—AVP) (From Dunn, et al. / Clin. Invest., 52:3212, 1973.)

The osmoreceptors respond to changes in extracellular fluid sodium concentration but very little to changes in potassium concentration and only slightly to changes in urea and glucose concentrations. Therefore, it must be emphasized again that, for all practical purposes, *the osmoreceptors are actually sodium concentration receptors*—hence the name *osmosodium receptors*.

The impulses from the osmoreceptors are transmitted from the supraoptic nuclei through the pituitary stalk into the posterior pituitary gland where they promote the release of antidiuretic hormone (ADH); further details of the secretion and release of ADH will be discussed in Chapter 75 in relation to the endocrinology of the pituitary gland; what is important here is that increased amounts of this hormone now act on the kidneys.

Earlier in the chapter it was pointed out that nervous signals from the volume receptors in the heart and also from the baroreceptors help control antidiuretic hormone secretion; this is in addition to the effect of changes in extracellular fluid osmolality in controlling its secretion. Figure 36–6 illustrates quantitatively the effects of these two separate stimuli on the concentration of ADH ("arginine vasopressin") in the plasma. The open circles show the effect of plasma osmolality, illustrating that an increase in plasma osmolality of about 6 per cent will increase the plasma ADH concentration to 10 to 20 times normal. The effect of decreased blood volume is illustrated by the solid circles, showing that relatively little increase in antidiuretic hormone concentration occurs until the blood volume is decreased more than 5 to 10 per cent. Since this much decrease occurs only a person bleeds severely, one can see that normally antidiuretic hormone secretion is controlled almost entirely by the changes in plasma osmolality.

Summary of the Antidiuretic Hormone Mechanism for Controlling Extracellular Fluid Osmolality and Extracellular Fluid Sodium Concentration. From these discussions, we can reiterate once again the importance of the antidiuretic hormone mechanism for controlling at the same time both extracellular fluid osmolality and extracellular fluid sodium concentration. That is, an increase in sodium concentration causes almost an exactly parallel increase in osmolality which in turn excites the osmoreceptors of the hypothalamus. These receptors then cause the secretion of antidiuretic hormone, which markedly increases the reabsorption of water in the renal tubules. Consequently, very little water is lost into the urine, but the urinary solutes continue to be lost. Therefore, the relative proportion of water in the extracellular fluid increases, whereas the proportion of solutes decreases. In this way, the sodium ion concentration of the extracellular fluid and the osmolality as well, decrease toward the normal level. This is a very powerful mechanism for controlling both the extracellular fluid osmolality and the extracellular fluid sodium concentration.

Function of the Antidiuretic Hormone Mechanism in Special Conditions

Water Diuresis. When a person drinks a large amount of water, a phenomenon called *water diuresis* occurs, and a typical record of this is shown in Figure 36–7. In this example, a man drank 1 liter of water, and approximately 45 minutes later his urine output had increased to eight times normal. It remained at this level for two hours—that is, until the osmolality of the extracellular fluid returned essentially to normal. The delay in onset of water diuresis is caused partly by dela

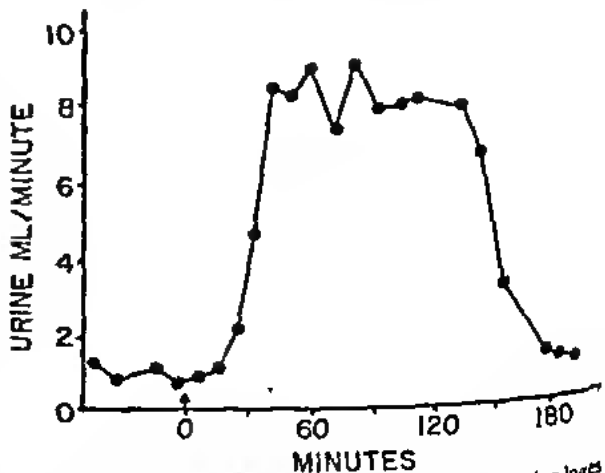


Figure 36–7. Water diuresis in a human being following ingestion of 1000 ml of water. (Redrawn from Smith: The Kidney: Structure and Functions in Health and Disease. Oxford University Press, 1951.)

in absorption of the water from the gastrointestinal tract but mainly by the time required for destruction of the antidiuretic hormone that had already been released by the pituitary gland prior to drinking the water.

Syndrome of Inappropriate ADH Secretion. Certain types of tumors, especially bronchogenic tumors of the lungs or tumors of the basal regions of the brain, often secrete antidiuretic hormone or a similar hormone. This condition is called the *syndrome of inappropriate ADH secretion*. This excess ADH causes only a slight increase in extracellular fluid volume, as pointed out earlier. Instead, its principal effect is *to decrease greatly the sodium concentration (and the osmolality) of the extracellular fluid*. The explanation of this effect is the following: The ADH at first causes a decrease in urine output and a simultaneous slight increase in blood volume. This in turn causes a slight rise in arterial pressure, which then causes a secondary *increase* in urinary output. Yet, the urine that is excreted is still tremendously concentrated because the antidiuretic hormone is still causing excessive reabsorption of water by the kidney tubules. Consequently, the kidneys excrete extreme amounts of sodium and other ions into the urine but keep the water in the extracellular fluid. Therefore, the sodium concentration becomes seriously reduced, sometimes falling from a normal value of 142 mEq/liter to as low as 110 to 120 mEq/liter. At values this low, patients frequently die sudden deaths because of coma and convulsions.

This disease is especially instructive because it illustrates the extreme importance of the antidiuretic hormone mechanism for control of sodium concentration and extracellular osmolality and yet its relatively mild effect on control of body fluid volume.

THIRST, AND ITS ROLE IN CONTROLLING SODIUM CONCENTRATION AND EXTRACELLULAR FLUID OSMOLALITY

The phenomenon of thirst is equally as important for regulating body water, sodium concentration, and osmolality as is the osmoreceptor-renal mechanism discussed above, because the amount of water in the body at any one time is determined by the balance between both *intake* and *output* of water. Thirst, the primary regulator of the intake of water, is defined as the *conscious desire for water*.

Neural Integration of Thirst—The "Thirst" Center

Referring again to Figure 36–5, one sees a small area located slightly anterior to the supraoptic nuclei in the lateral preoptic area of the hypothalamus, which is called the *thirst center*. Electrical stimulation of this center by implanted electrodes causes an animal to begin drinking within seconds and to continue drinking until the electrical stimulus is stopped. Also, injection of hypertonic salt solutions into the area, which causes osmosis of water out of the neuronal cells, thus shrinking the cells, also causes drinking. Thus, the neuronal cells of the thirst center function in almost the same way as the osmoreceptors of the supraoptic nuclei.

An increase in osmotic pressure of the cerebrospinal fluid in the third ventricle has essentially the same effect of promoting thirst as an increase in the osmotic pressure of the circulating extracellular fluid. Therefore, it has been suggested that the primary site for detecting changes in osmolality might be at the surface of the third ventricle and that appropriate signals are then transmitted to the thirst center to cause drinking. However, the importance of this is still debated.

Basic Stimulus for Exciting the Thirst Center—Intracellular Dehydration. Any factor that will cause *intracellular dehydration* will in general cause the sensation of thirst. The most common cause of this is increased osmolar concentration of the extracellular fluid, especially increased sodium concentration, which causes osmosis of fluid from the neuronal cells of the thirst center. However, another important cause is excessive loss of potassium from the body, which reduces the intracellular potassium of the thirst cells and therefore decreases their volume.

Other Stimuli That Lead to Thirst

Stimulation of Thirst by Angiotensin. Excess angiotensin II in the circulating body fluids often increases the rate of drinking by animals as much as two- to threefold. Therefore, almost any circulatory condition that leads to increased production of angiotensin II will also lead to thirst and drinking. In addition, injection of angiotensin into the third ventricle has an especially strong effect on promoting drinking, but this can be prevented by destroying the anteroventral surface of the third ventricle. Therefore, it has been suggested that the major site of action of angiotensin in causing drinking is not on the thirst center itself but instead on a specialized center located immediately beneath the surface of the third ventricle, and that signals are conducted from this area to the thirst center to increase drinking.

Stimulation of Thirst by Hemorrhage and Low Cardiac Output. A small amount of hemorrhage ordinarily does not cause thirst, but loss of as much as 10 per cent of the blood volume, which also reduces the cardiac output significantly, usually does lead to thirst. Also, patients with low cardiac output resulting from cardiac failure frequently develop intense thirst. In both of these conditions excessive amounts of angiotensin II are formed, and it is possible that this is the stimulus for the thirst.

Dryness of the Mouth. Almost everyone is aware that a dry mouth is often associated with thirst. The probable explanation for this is that the same factors that cause intracellular dehydration and therefore stimulate the thirst center also cause a dry mouth. Therefore, we have come to associate a dry mouth with the thirst sensation. However, in opposition to this concept is the fact that in animal experiments in which a dry mouth has been achieved by blocking secretion by the salivary glands, the animals do not drink excessively except under one condition: when they are eating food. This seems to result from the need for water to lubricate the

swallowing of the food and not from thirst. The same effect occurs in human beings whose salivary glands do not secrete saliva.

Temporary Relief of Thirst Caused by the Act of Drinking

A thirsty person receives relief from thirst immediately after drinking water, even before the water has been absorbed from the gastrointestinal tract. In fact, in persons who have an esophageal opening to the exterior so that the water is lost to the outside and never goes into the gastrointestinal tract, partial relief of thirst still occurs following the act of drinking, but this relief is only temporary, and the thirst returns after 15 or more minutes. If the water does enter the stomach, distension of the stomach and other portions of the upper gastrointestinal tract provides still further temporary relief from thirst. For instance, simple inflation of a balloon in the stomach can relieve thirst for 5 to 30 minutes.

One might wonder what the value of this temporary relief from thirst could be, but there is good reason for its occurrence. After a person has drunk water, as long as one half to one hour may be required for all the water to be absorbed and distributed throughout the body. Were the thirst sensation not temporarily relieved after drinking water, the person would continue to drink more and more. When all this water should finally become absorbed, the body fluids would be far more diluted than normal, and an abnormal condition opposite to that which the person was attempting to correct would have been created. It is well known that a thirsty animal almost never drinks more than the amount of water needed to relieve its state of dehydration. Indeed, it is uncanny that the animal usually drinks almost exactly the right amount.

Role of Thirst in Controlling Extracellular Fluid Osmolality and Sodium Concentration

Threshold for Drinking—The Tripping Mechanism. The kidneys are continually excreting fluid; also water is lost by evaporation from the skin and lungs. Therefore, a person is continually being dehydrated, causing the volume of extracellular fluid to decrease and its concentration of sodium and other osmolar elements to rise. When the sodium concentration rises approximately 2 mEq/liter above normal (or the osmolality rises approximately 4 mOsm/liter above normal) the drinking mechanism becomes "tripped"; that is, the person then reaches a level of thirst that is strong enough to activate the necessary motor effort to cause drinking. This is called the *threshold for drinking*. The person ordinarily drinks ly the required amount of fluid to bring the extracellular fluids back to normal—that is, it state of *satiety*. Then the process of dehydrati and sodium concentration begins again, and af a period of time the drinking act is tripped aga the process continuing on and on indefinitely.

In this way, both the sodium concentration a the osmolality of the extracellular fluid are v precisely controlled.

COMBINED ROLES OF THE ANTIDIURETIC AND THIRST MECHANISMS FOR CONTROL OF EXTRACELLULAR FLUID SODIUM CONCENTRATION AND OSMOLALITY

When either the antidiuretic hormone mec nism or the thirst mechanism fails, the ot ordinarily can still control both sodium concent tion and extracellular fluid osmolality with reas able effectiveness. On the other hand, if both them fail simultaneously, neither sodium nor molality is then adequately controlled.

Figure 36–8 illustrates dramatically the over capability of the ADH-thirst system to cont extracellular fluid sodium concentration. This ure demonstrates the ability of the same anim to control its extracellular fluid sodium concent tion in two different conditions. (1) in the norm state, and (2) after both the antidiuretic hormo and thirst mechanisms had been blocked. N that in the normal animal (the solid curve sixfold increase in sodium intake caused the dium concentration to change only two thirds o per cent (from 142 mEq/liter to 143 mEq/liter an excellent degree of sodium concentration c

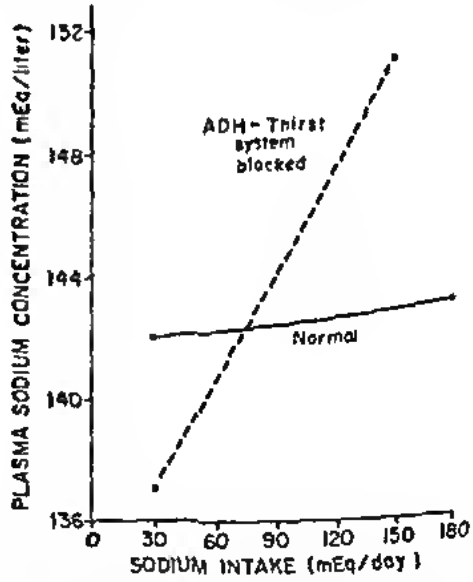


Figure 36–8. Effect on the extracellular fluid sodium concent tion in dogs caused by tremendous changes in sodium intake under normal conditions, and (2) after the antidiuretic hormo and thirst feedback systems had been blocked. This figure sho lack of sodium ion control in the absence of these syste (Courtesy of Dr David B Young)

rol. Now note the dashed curve of the figure, which shows the change in sodium concentration when the ADH-thirst system was blocked. In this case, the sodium concentration increased 10 per cent with only a fivefold increase in sodium intake a change in sodium concentration from 137 mEq/liter to 151 mEq/liter), which is an extreme hange in sodium concentration when one realizes hat the normal sodium concentration rarely rises or falls more than 1 per cent from day to day.

Therefore, the major feedback mechanism for ontrol of sodium concentration (and also for extracellular osmolality) is the ADH-thirst mechanism. In the absence of this mechanism there is o feedback mechanism that will cause the body o increase water ingestion or water conservation by the kidneys when excess sodium enters the body. Therefore, the sodium concentration simply increases.

EFFECT OF ALDOSTERONE ON SODIUM CONCENTRATION

A second hormonal system that plays a *small* role in controlling extracellular fluid sodium concentration is the aldosterone feedback system. Figure 36–9 illustrates how slight the effect of the aldosterone system is in controlling plasma sodium concentration. This figure shows the effect on sodium concentration of more than a sixfold increase in sodium intake in the same dog (a) under normal conditions, and (b) after the aldosterone system had been blocked—that is, the adrenal glands had been removed and the animals were infused at a constant rate of aldosterone that could neither change upward nor downward. Note that in the normal state the sodium concentration changed exactly 1 per cent, while when the aldosterone system was blocked it changed exactly 2 per cent. In other words, even without a functional aldosterone feedback system (because the aldosterone could neither change upward nor downward) sodium concentration was still very well regulated.

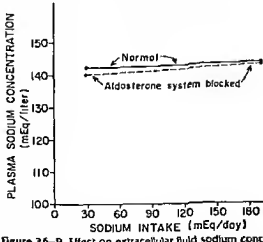


Figure 36–9. Effect on extracellular fluid sodium concentration in dogs caused by tremendous changes in sodium intake (1) under normal conditions, and (2) after the aldosterone feedback system had been blocked. Note that sodium is exceedingly well controlled with or without aldosterone feedback control. (Courtesy of Dr. David B Young)

Because of the great effect that aldosterone has on tubular sodium reabsorption, this lack of importance of aldosterone for regulation of sodium concentration seems to be a paradox, but it results from the following simple effect: When the aldosterone causes increased sodium reabsorption from the tubules, as was discussed in the previous chapter, this causes a simultaneous reabsorption of water and an increase in extracellular fluid volume. An increase of only a few per cent in the extracellular fluid volume eventually leads to an increase in arterial pressure, and the increase in arterial pressure to increased glomerular filtration rate, a well known effect in the presence of excess aldosterone. The rapid flow of filtrate down the tubular system then compensates for the excessive reabsorptive effect of the aldosterone and thereby almost completely nullifies the effect of aldosterone on extracellular fluid sodium concentration.

Furthermore, as was explained above, the ADH-thirst system is an extremely powerful controller of sodium concentration—rouch more powerful than the aldosterone feedback system—so that the ADH-thirst system greatly overshadows the aldosterone system for sodium control under normal conditions. Indeed, even in patients who have primary aldosteronism (these patients secrete tremendous quantities of aldosterone) the sodium concentration still rises only 2 to 3 mEq/liter above normal

CONTROL OF SODIUM INTAKE—APPETITE AND CRAVING FOR SALT

Maintenance of normal extracellular sodium requires not only the control of sodium excretion but also the control of sodium intake. Unfortunately, we know very little about this except that salt-depleted persons (or persons who have lost blood) develop a desire for salt; as an example, this occurs in persons who have *Addison's disease*, a condition in which the adrenal cortices no longer secrete aldosterone so that the salt stores of the body become depleted. The salt-depleted person craves and eats naturally salty foods. Likewise, it is well known that animals living in areas far removed from the seashore actively search out "salt licks." This craving for salt is analogous to thirst, and it is also analogous to appetite for other types of foods, which is still another homeostatic mechanism that will be discussed in Chapter 73.

REGULATION OF POTASSIUM ION CONCENTRATION

It is especially important to control the extracellular fluid potassium concentration because very slight changes in concentration often can alter nervous and cardiac functions seriously, as discussed in Chapter 10 for peripheral nerves, in Chapter 46 for the central nervous system, and Chapter 13 for the heart. The normal concentration is about 4.5 mEq/liter, and this rarely rises or falls more than ± 0.3 mEq.

Two major factors play important roles in controlling the potassium ion concentration: direct effect of increased extracellular

sium concentration in causing increased secretion of potassium into the tubules, and (2) an effect of aldosterone in increasing potassium secretion.

The Direct Effect of Extracellular Fluid Potassium Concentration on Potassium Secretion into the Tubules

Potassium is secreted into the tubular fluid in the late distal tubules and cortical collecting tubules. This secretion is coupled with the active transport of sodium by the basolateral membranes of the tubular epithelial cells because potassium ions are transported in the opposite direction. Also, sodium transport from the interior of the epithelial cells into the interstitial fluids decreases the intracellular electrical potential to about −70 millivolts, and this negativity causes additional diffusion of the positively charged potassium ions from the interstitial fluid to the interior of the epithelial cells. Then, because the luminal border of the epithelial cells is reasonably permeable to potassium, the excess potassium diffuses the rest of the way into the tubular fluid. Since it is partly the electrical potential that causes the potassium movement, the rate of transport is determined to a great extent by the concentration of the potassium ions in the interstitial fluid of the kidneys, which is part of the body's extracellular fluid. Consequently, as the extracellular fluid potassium concentration increases, the rate of potassium secretion into the tubules also increases markedly. This is illustrated by the lowest curve in Figure 36–10, which shows that even a slight increase of the potassium ion concentration, from 3.2 to 4.4 mEq/liter, increases urinary potassium excretion almost eightfold. This obviously functions as an automatic means for reducing the extracellular fluid potassium ion concentration back toward normal.

Effect of Aldosterone on Tubular Secretion of Potassium Ions

The hormone aldosterone not only increases sodium reabsorption by the late distal tubules and collecting tubules, but at the same time it also increases potassium ion secretion into these tubules. This is illustrated by the top curve of Figure 36–10, showing that an increase of plasma aldosterone secretion to five times normal causes an increase in urinary potassium excretion somewhat over three times.

It is believed that the aldosterone increases potassium secretion secondarily to its effect on sodium reabsorption. That is, the greater the sodium reabsorption, the more rapid is the transport of potassium ions into the tubules.

Effect of Potassium Ion Concentration on Rate of Aldosterone Secretion. In any properly functioning feedback control system, the factor

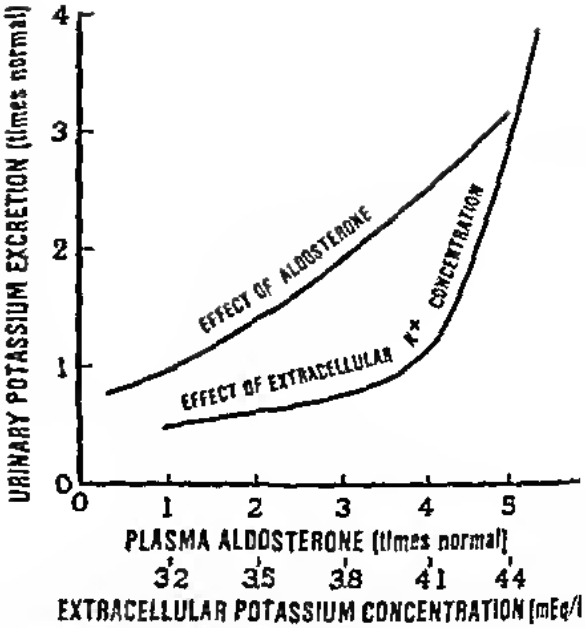


Figure 36–10. Effect of (a) plasma aldosterone concentr and (b) extracellular potassium ion concentration on the rat urinary potassium excretion. These are the two most impor factors regulating the rate of potassium excretion in the u (Drawn from data in Young: *Am. J. Physiol.* 244:F28, 1983.)

that is controlled almost invariably has a feedb effect for control of the controller. In the case the oldosterone-potassium control system, the r of aldosterone secretion by the zona glomerul cells of the adrenal cortex is controlled v strongly by the extracellular fluid potassium c centration. Figure 36–11 illustrates this, show that an increase in the potassium concentratio slightly more than 2 mEq/liter can increase plasma aldosterone concentration from almost 0 as high as 60 nanograms per 100 milliliters plasma, a concentration almost ten times norm

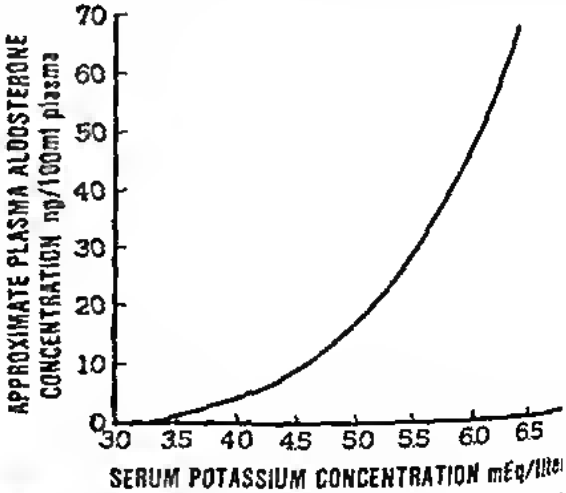


Figure 36–11. Effect on extracellular fluid aldosterone conc tration of potassium ion concentration changes. Note the extre change in aldosterone concentration for very minute change potassium concentration.

Putting the effects illustrated in Figure 36–11 gether with the fact that aldosterone greatly creases renal excretion of potassium (as shown Figure 36–10), one can construct a very simple ystem for negative feedback control of potassium ncentration, illustrated in Figure 36–12. That , an increase in potassium concentration causes a increase in aldosterone concentration in the rculating blood (Block 1). The increase in aldo- erone concentration then causes a marked in- rease in potassium excretion by the kidneys Block 2). The increased potassium excretion then ecreases the extracellular fluid potassium concen- ration back toward normal (Blocks 3 and 4).

Importance of the Aldosterone Feedback ystem for Control of Potassium Concentra- ion. Without a functioning aldosterone feedback ystem, an animal can easily die from hyperpotas- emia.

Figure 36–13 illustrates the potent effect of the ... feedback system in controlling potas- ium concentration. In the experiment of this fig- ire, a series of dogs was subjected to an almost evenfold increase in potassium intake in two ifferent states: (1) tha normal state, and (2) after ha aldosterone feedback system had been blocked y removing tha adrenal glands and tha animals iven a fixed rate of aldosterone infusion.

Note that in tha normal animal the sevenfold ncrease in potassium intaka caused an increasa in plasma potassium concentration of only 2.4 per cent—from a concentration of 4.2 mEq/liter to 4.3 mEq/liter. Thus, when tha aldosterona feedback system was functioning normally, the potassium concentration remained very precisely controlled despite the tremendous change in potassium in- take.

On tha other hand, the dashed curve in tha figure shows the effect after the aldosterone system had been blocked. Note that the same increase in potassium intake now caused a 26 per cent in- crease in potassium concentration! Thus, the con- trol of potassium concentration in the normal an- imals was many times as effective as in the animals without an aldosterone feedback mecha- nism.

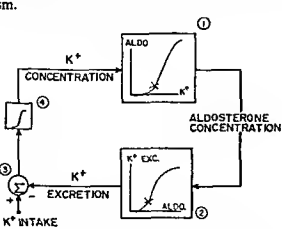


Figure 36–12. Simplified schema of the aldosterone system for control of extracellular fluid potassium concentration.

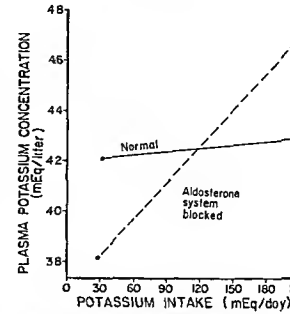


Figure 36–13. Effect on extracellular fluid potassium concentration of tremendous changes in potassium intake (1) under normal conditions, and (2) after the aldosterone feedback system had been blocked. This figure demonstrates that potassium concentration is very poorly controlled after block of the aldosterone system. (Courtesy of Dr David B. Young.)

Effect of Primary Aldosteronism and Addison's Disease on Extracellular Fluid Potassium Concentration. *Primary aldosteronism* is caused by a tumor of the zona glomerulosa of one of the adreoal glands, the tumor secreting tremendous quantities of aldosterone. One of the most important effects of this disease is a severe decrease in extracellular fluid potassium concentration, so much so that many of these patients experience paralysis caused by failura of nerve transmission, es was explained in Chapter 10.

Conversely, in *Addison's disease* the adrenal glands have been destroyed, and the extracellular fluid potassium concentratioo frequently rises to as high as double normal. This is often the cause of death in these patients, resulting in cardiac arrest.

OTHER FACTORS THAT AFFECT POTASSIUM ION CONCENTRATION

Other factors that affect the extracellular fluid potassium ion concentration include (1) changes in the hydrogen ion concentration (because hydrogen competes with potassium for secretion by the kidney tubules), and (2) the level of sodium intake (because sodium is reciprocally transported through the late distal tubule, collecting tubule, and collecting duct epithelium in exchange for potassium). However, the effects of both these factors are much less than the effects of extracellular potassium concentration and of aldosterone.

CONTROL OF THE EXTRACELLULAR CONCENTRATIONS OF OTHER IONS

Regulation of Calcium Ion Concentration

The role of calcium in the body and control of its concentration in the extracellular fluid will be discussed in detail in Chapter 79 in relation to the

of parathyroid hormone, calcitonin, and bone However, briefly, it is the following

The *day-by-day* calcium ion concentration remains within a few per cent of 2.4 mEq/liter and is *controlled principally by the effect of parathyroid hormone on bone reabsorption.* When the extracellular fluid concentration of calcium falls too low, the parathyroid glands are directly stimulated to promote increased secretion of parathyroid hormone This hormone in turn acts directly on the bones to increase the reabsorption of bone salts, thus releasing large amounts of calcium into the extracellular fluid and elevating the calcium level back to normal On the other hand, when the calcium concentration becomes too great, parathyroid hormone secretion becomes depressed so that almost no bone reabsorption then occurs. Yet, the osteoblastic system for forming new bone does continue to deposit calcium, thus removing calcium from the extracellular fluid, in this way reducing the calcium ion concentration back to normal

However, the bones are not an inexhaustible supply of calcium, and eventually the bones will run out of calcium Therefore, long-term control of calcium ion concentration results from the effect of parathyroid hormone on reabsorption of calcium from the kidney tubules and absorption of calcium from the gut through the gastrointestinal mucosa, both of which effects are markedly increased by parathyroid hormone

The effect of parathyroid hormone on calcium handling by the kidneys closely parallels the effect of aldosterone on sodium handling by the kidneys. That is, even in tha absence of parathyroid hormone, much of tha calcium is reabsorbed from the tubular fluid in the proximal tubules, the loop of Henle, and the diluting segment of the distal tubules, but about 10 per cent of the filtered load of calcium still remains to enter the late distal tubules Then, if large amounts of parathyroid hormone are present in the body fluids, essentially all the remaining calcium will be reabsorbed from the late distal tubules, the collecting tubules, and collecting ducts, thus conserving the calcium in the body So the principal mechanism for *long-term control of calcium ion concentration* is low calcium in the extracellular fluid leads to parathyroid hormone secretion and the parathyroid hormone then promotes intense *conservation of calcium by the kidneys and also greatly increased calcium absorption from the gastrointestinal tract.* These effects will be discussed in more detail in Chapter 79.

Regulation of Magnesium Ion Concentration

Much less is known about the regulation of magnesium ion concentration than of calcium ion concentration Magnesium ions are reabsorbed by all portions of the renal tubules However, the magnesium ion itself directly affects the tubular epithelial cells by decreasing this reabsorption, especially in the thick segment of the ascending limb of the loop of Henle Therefore, when the extracellular fluid magnesium concentration is high, excess amounts of magnesium are excreted, and, conversely, when the magnesium concentration is low, magnesium is conserved.

Regulation of Phosphate Concentration

Phosphate concentration is regulated primarily by an *overflow* mechanism, which can be explained as follows: The renal tubules have a normal transport maximum for absorbing phosphate of 0.1 millimole per minut When less than this "load" of phosphate is present the glomerular filtrate, all of it is reabsorbed. Wh more than this amount is present, the excess is excre Therefore, normally phosphate ion spills into the uri when its concentration in the extracellular fluid is abo the threshold value of approximately 0 8 millimole/lit Since most people ingest large quantities of phosph day in and day out, either in milk or in meat, t concentration of phosphate is usually maintained at level of about 1 0 millimole/liter, a level at which the is continual overflow of excess phosphate into the uri

Role of Parathyroid Hormone in Phosphate I Regulation. Parathyroid hormone, which plays a ma cole in regulation of calcium ion concentration, as e plained above, also affects phosphate ion concentrati in two different ways. First, parathyroid hormone p motes bone reabsorption, thereby dumping large qua tities of phosphate ions into the extracellular fluid fr the bone salts Second, this hormone decreases t transport maximum for phosphate by the renal tubul so that a greater proportion of the tubular phosphate lost in the urine The combination of these fact causes, at least acutely, marked loss of phosphates the urine The interrelationships between phosph calcium, and parathyroid hormone control will be d cussed in Chapter 79

Regulation of Other Negative Ions. The two m important negative ions in the extracellular fluid a *chloride ions* and *bicarbonate ions* The concentratio of these are normally 103 mEq/liter and 28 mEq/li respectively When the positive ions in the extracellul fluid (mainly the sodium ions) increase, the total qua tity of the negative ions increases at the same tu because they are literally pulled through the tubul epithelium by the electrical charges of tha positive io However, the relative proportions of the chloride a bicarbonate ions are determined mainly by the sa control mechanisms that control the hydrogen ion co centration We will discuss these mechanisms in t following chapter Other important negative ions in t body fluid include sulfates, nitrates, urates, lactat and the amino acids Essentially all of these, like phc phate, have definite absorptive transport maximum When the concentration of each is below its respecti threshold, it is conserved in the extracellular fluid, b when above this threshold the excess spills into t urine. Thus, the concentrations of most of these negati ions are regulated by the overflow mechanism in t same way that phosphate ion concentration is regulate

REFERENCES

Altura, B M., and Altura, B T Magnesium, electrolyte transport a coronary vascular tone *Drugs, 28*(Suppl 1) 120, 1984

Andersson, B Regulation of body fluids *Annu Rev Physiol*, 39 1 1977

Andersson, B Regulation of water intake *Physiol. Rev*, 58 582 197

Andersson, B, and Rundgren, M. Thirst and its disorders *Annu. R Med*, 33 231, 1982

Angus, Z S., *et al*, PTH, calcitonin, cyclic nucleotides, and the kidn *Annu Rev Physiol*, 43 583 1981

Baer, P G, and McGiff, J C Hormonal systems and renal hemod namics *Annu Rev Physiol*, 42 589, 1980

Blaustein, M P, and Hamlyn, J M Sodium transport inhibition, c calcium, and hypertension The natriuretic hormone/Na+-Ca2+ e change hypertension hypothesis *Am. J Med*, 77 45, 1984

Boron, W F, and Sackin, H Measurement of intracellular ion composition and activities in renal tubules *Annu Rev Physi* 45 483, 1983

Brownstein, M. J.: Biosynthesis of vasopressin and oxytocin *Annu. Rev Physiol*, *45* 129, 1983
Buckalew, V M., Jr, and Gruber, K A Natriuretic hormone *Annu Rev Physiol*, *46*:343, 1984
Crawford, M P, *et al* · Renal servocontrol of arterial blood pressure *J Appl. Physiol*, *22* 139, 1967.
Cross, B. A, and Wakerley, J. B: The neurohypophysis. *Int Rev Physiol*, *16* 1, 1977.
Currie, M C., and Needleman, P Renal arachidonic acid metabolism, *Annu. Rev. Physiol*, *46* 327, 1984
Dawson, K G Endocrine physiology of electrolyte metabolism *Drugs*, 28(Suppl 1) 98, 1984.
Dennis, V W., *et al*. Renal handling of phosphate and calcium *Annu Rev. Physiol*, *41* 257, 1979
Doris, P. A. Vasopressin and central integrative processes. *Neuroendocrinology*, 38 75, 1984
M J. (ed) Prostaglandins and the Kidney Biochemistry. Physiology, Pharmacology, and Clinical Applications New York, Plenum Publishing Corp., 1983
Earley, L. E, and Schrier, R W Intrarenal control of sodium excretion by hemodynamic and physical factors. *In* Orloff, F, and Berliner, R. W (eds) Handbook of Physiology Sec 8 Baltimore, Williams & Wilkins, 1973, p 721
Ehrlich, E N Adrenocortical regulation of salt and water metabolism Physiology, pathophysiology, and clinical syndromes *In* DeGroot, L. J, *et al*. (eds) Endocrinology Vol 3 New York, Grune & Stratton, 1979, p. 1883
Farquhar, M. D, and Park, C S Steroid hormones and the kidney *Annu. Rev Physiol*, *43* 637, 1981
Fitzsimons, J T · Thirst *Physiol Rev*. *52* 468, 1972
Flatman, P W Magnesium transport across cell membranes *J Membr Biol*, *60* 1, 1984
Fregly, M. J., and Rowland, N E Role of renin-angiotensin aldosterone system in NaCl appetite of rats *Am J Physiol*, *248* R1, 1985
Gauer, D H, and Henry, J P Neurohormonal control of plasma volume. *Int Rev Physiol*, *9* 145, 1976
Gibbons, G H., *et al* Interaction of signals influencing renin release *Annu Rev Physiol*, *46* 231, 1984
Goetz, K. L., *et al*. Atrial receptore and renal function *Physiol Rev*, *55* 157, 1975
Gottschalk, C W Renal nerves and sodium excretion *Annu Rev Physiol*, *41*:229, 1979
Guignard, J P., and Filloux, B Studies on compensatory adaptation of renal functions *Yale J Biol Med*, *51*(3) 247, 1978
Guyton, A. C, *et al* Theory for renal autoregulation by feedback at the juxtaglomerular apparatus *Circ Res*, *14* 187, 1964
Guyton, A C, *et al* Dynamics and Control of the Body Fluids. Philadelphia, W B Saunders Co, 1975
Guyton, A C, *et al* A system analysis of volume regulation Alfred Benzon Symposium XI Munksgaard, 1978
Hall, J E, *et al* Control of glomerular filtration rate by renin-angiotensin system. *Am J Physiol*., *233* F366, 1977.
Hall, J E, *et al*., Dissociation of renal blood flow and filtration rate autoregulation by renin depletion *Am J Physiol*, *232* F215, 1977
Hall, J E, *et al*. Intrarenal control of electrolyte excretion by angiotensin II *Am J Physiol*, *232* F538, 1977
Hall, J E, *et al*, Intrarenal role of angiotensin II and [des-Asp¹] angiotensin II. *Am J Physiol*, *236* F252, 1979
Handler, J S, and Orloff, J Antidiuretic hormone *Annu Rev Physiol*, *43* 611, 1981
Harrison, M H. Effects of thermal stress and exercise on blood volume in humans *Physiol Rev*, *65* 149, 1985
Heyward, J N Neural control of the posterior pituitary *Annu Rev Physiol*, *37* 191, 1975
Insel, P A, and Snavely, M. D Catecholamines and the kidney Receptors and renal function *Annu Rev Physiol*, *43* 625, 1981
Jackson, T. E, *et al* Transient response of glomerular filtration rate and renal blood flow to step changes in arterial pressure *Am J Physiol*, *233* F395, 1977
Katz, A I, and Lindheimer, M D Actions of hormones on the kidney *Annu Rev Physiol*, *39* 97, 1977
Knox, F G, and Diaz Buxo, J A The hormonal control of sodium excretion *Int Rev Physiol*, *16* 173, 1977
Margolius, H S The kallikrein-kinin system of the kidney *Annu Rev Physiol*, *46* 309, 1984
Maxwell, M H, and Kleeman, C R. (eds.) Clinical Disorders of Fluid and Electrolyte Metabolism, 3rd Ed New York, McGraw-Hill, 1979
Menninger, R P Current concepts of volume receptor regulation of vasopressin release *Fed Proc*., *41* 55, 1985
Miller, M Assessment of hormonal disorders of water metabolism *Clin Lab Med*, *4* 729, 1984
Naftalin, R J The thermostatics and thermodynamics of cotransport *Biochem Biophys Acta*, *778* 155, 1984
Nagelatti, A, and Malik, K U The renal kallikrein-kinin and prostaglandin systems interaction *Annu Rev Physiol*, *43* 597, 1981
Navar, L G, *et al* Effect of alterations in plasma osmolality on renal blood flow autoregulation *Am. J Physiol*, *211* 1387, 1966
Navar, L. G, *et al* Distal tubular feedback control of renal hemodynamics and autoregulation *Annu Rev Physiol*, *42* 557, 1980
O Connor, W J Normal Renal Function The Excretion of Water and Electrolytes Derived from Food and Drink New York, Oxford University Press, 1982
Robertson, G L. Vasopressin in osmotic regulation in man *Annu Rev Med*, *25* 315, 1974
Sawyer, W H, and Manning, M The use of antagonists of vasopressin in studies of its physiological functions *Fed Proc*, *44* 78, 1985
Schmidt-Nielsen, B, and Lawe, D F Invertebrate mechanisms for diluting and concentrating the urine *Annu Rev Physiol*, *25* 631, 1963
Sklar, A H, and Schrier, R. W Central nervous system mediators of vasopressin release *Physiol Rev*, *63* 1243, 1983
Smith, M J, Jr, *et al* Acute and chronic effects of vasopressin on blood pressure, electrolytes, and fluid volumes *Am. J Physiol*, *237*(3) F232, 1979
Stein, J H., and Reineck, H J Effect of alterations in extracellular fluid volume on segmental sodium transport *Physiol Rev*, *55* 127, 1975
Sutton, R A Diuretics and calcium metabolism *Am J Kidney Dis*, *5* 4, 1985
Swanson, L W, and Sawchenko, P E Hypothalamic integration Organization of the paraventricular and supraoptic nuclei *Annu Rev Neurosci*, *6* 269, 1983
Trippodo, N C, *et al* Intrarenal role of angiotensin II in controlling sodium excretion during dehydration in dogs *Clin Sci Molec Med*, *52* 545, 1977
Vallan, H Renal Dysfunction. Mechanisms Involved in Fluid and Solute Imbalance Boston, Little, Brown, 1979
Verney, E B Absorption and excretion of water Antidiuretic hormone *Lancet*, *2* 739, 1946
Wolf, A V Thirst. Physiology of the Urge to Drink and Problems of Water Lack Springfield, Ill, Charles C Thomas, 1958
Young, D B, *et al* Effectiveness of the aldosterone-sodium and potassium feedback control system *Am J Physiol*, *231* 945, 1976
Young, D B, *et al* Control of extracellular sodium concentration by antidiuretic hormone-thirst feedback mechanism. *Am J Physiol*, *232* R145, 1977
Zerbe, R *et al*. Vasopressin function in the syndrome of inappropriate antidiuresis *Annu Rev Med*, *31* 315, 1980
Zusman, R. M Prostaglandins and water excretion *Annu Rev Med*, *32* 359, 1981

37

Regulation of Acid-Base Balance

When one speaks of the regulation of acid-base balance, regulation of hydrogen ion concentration in the body fluids is actually meant. The hydrogen ion concentration in different solutions can vary from less than 10^{-14} equivalents per liter to higher than 10^{0}, which means a total variation of more than a quadrillionfold. On a logarithmic basis, the hydrogen ion concentration in the human body is approximately midway between these two extremes.

Only slight changes in hydrogen ion concentration from the normal value can cause marked alterations in the rates of chemical reactions in the cells, some being depressed and others accelerated. For this reason the regulation of hydrogen ion concentration is one of the most important aspects of homeostasis. Later in the chapter the overall effects of high hydrogen ion concentration (acidosis) and low hydrogen ion concentration (alkalosis) are discussed. In general, when people become acidotic they are likely to die in coma, and when they become alkalotic they may die of tetany or convulsions.

Acids and Bases—Their Definitions and Meanings

An acid is defined as a molecule or an ion that can function as a *proton donor*. And a base is defined as a molecule or an ion that can function as a *proton acceptor*. If one will remember that a proton is actually a hydrogen ion, one can restate these definitions in the following ways.

An *acid* is a molecule or an ion that can contribute a hydrogen ion to a solution. Thus, HCl ionizes in water to form hydrogen ions and chloride ions and therefore is the acid called hydrochloric acid. Likewise, H_2CO_3 ionizes in water to form H^+ and HCO_3^-, thus also contributing hydrogen ions to the solution, and is the acid known as carbonic acid. Other important acids in the body are acetic acid, sodium acid phosphate, uric acid, acetoacetic acid, and so forth.

A *base* is a molecule or an ion that will combine with hydrogen ions to remove them from a solution. For instance, the bicarbonate ion, HCO_3^-, is a base because it can combine with hydrogen ions to form H_2CO_3.

HPO_4^{--} is a base because it can combine with hydrogen ions to form $H_2PO_4^-$. The proteins of the body also function as bases because certain of the amino acids in the protein molecules function as negative ions that bind readily with excess hydrogen ions. In fact, the hemoglobin in the red blood cells and the other proteins in the other cells of the body are among the most important of the body's bases. Commonly, molecules such as sodium bicarbonate and sodium phosphate are called bases as well. However, the negative ions of these molecules are actually the real bases, according to the most-used definition of the word "base."

Finally, one needs to understand the relationship between a "base" and an "alkali." An alkali is the combination of one of the alkaline metals—sodium, potassium, or so forth—with a highly basic ion such as the hydroxyl ion (OH^-). The basic portions of these molecules react vigorously with hydrogen ions to remove them from the solutions and therefore are typical bases. Since alkalis are well known, the term alkali is frequently used synonymously with the term base. And for similar reasons, the term *alkalosis* is used to mean the opposite of acidosis, meaning excess removal of hydrogen ions from the solution, in contrast to excess addition of hydrogen ions in the state of acidosis.

Strong and Weak Acids and Bases. A strong acid is one in which the molecule or ion that acts as a hydrogen ion donor has a very strong tendency to dissociate into ions and therefore to discharge its hydrogen ion into the solution. A typical example is hydrochloric acid. On the other hand, acids that release hydrogen ions with far less vigor are called weak acids. Examples are carbonic acid and sodium acid phosphate. A strong base is one that reacts very powerfully with hydrogen ions and therefore removes these with extreme avidity from the solution. A typical example is the hydroxyl ion (OH^-). A typical weak base is the bicarbonate ion (HCO_3^-), for it binds much more weakly with hydrogen ions.

Most of the acids and bases that are concerned with normal regulation of acid-base balance in the body are weak acids and bases, the most important of which are carbonic acid and bicarbonate base.

Hydrogen Ion Concentration and pH of Normal Body Fluids and in Acidosis and Alkalosis. The hydrogen ion concentration in the extracellular fluid is normally regulated at a constant value of approximately 4×10^{-8} Eq/liter; this value can vary from as low as 1.0×10^{-8} to as high as 1.6×10^{-7} without causing death.

From these values, it is already apparent that

xpressing hydrogen ion concentration in terms of ctual concentrations is a cumbersome procedure. herefore, the symbol *pH* has come into usage for xpressing the concentration, and pH is related to ctual hydrogen ion concentration by the following ormula (when H^+ concentration is expressed in quivalents per liter):

$$pH = \log \frac{1}{H^+ \text{ conc.}} = -\log H^+ \text{ conc.} \quad (1)$$

Note from this formula that a low pH corresponds to a high hydrogen ion concentration, which s called acidosis; and, conversely, a high pH corresponds to a low hydrogen ion concentration, which is called *alkalosis*.

The normal pH of arterial blood is 7.4, whereas he pH of venous blood and of interstitial fluids is bout 7.35 because of extra quantities of carbon lioxide that form carbonic acid in these fluids.

Since the normal pH of the arterial blood is 7.4, a person is considered to have acidosis whenever pH is below this value and to have alkalosis when it rises above 7.4. The lower limit at which a person can live more than a few hours is about 6.8, and the upper limit approximately 8.0.

Intracellular pH. On the basis of indirect measurements, it has been found that the intracellular pH usually ranges between 6.0 and 7.4 in different cells, perhaps averaging about 7.0. A *rapid rate of metabolism* in cells increases the rate of acid formation, especially carbonic acid (H_2CO_3), and consequently decreases pH. Also, *poor blood flow* to any tissue causes acid accumulation and a decrease in pH.

DEFENSE AGAINST CHANGES IN HYDROGEN ION CONCENTRATION

To prevent acidosis or alkalosis, several special control systems are available: (1) All the body fluids are supplied with acid-base *buffer systems* that immediately combine with any acid or alkali and thereby prevent excessive changes in hydrogen ion concentration. (2) If the hydrogen ion concentration does change measurably, the *respiratory center is immediately stimulated* to alter the rate of breathing. As a result, the rate of carbon dioxide removal from the body fluids is automatically changed, and, for reasons that will be presented later, this causes the hydrogen ion concentration to return toward normal. (3) When the hydrogen ion concentration changes from normal, *the kidneys excrete either an acid or alkaline urine*, thereby also helping readjust the hydrogen ion concentration of the body fluids back to normal.

The buffer systems can act within a fraction of a second to prevent excessive changes in hydrogen ion concentration. On the other hand, it takes 1 to 15 minutes for the respiratory system to readjust the hydrogen ion concentration after a sudden change has occurred. Finally, the kidneys, though providing the most powerful of all the acid-base regulatory systems, require many minutes to several days to readjust the hydrogen ion concentration.

FUNCTION OF ACID-BASE BUFFERS

An acid-base buffer is a solution of two or more chemical compounds that prevents marked changes in hydrogen ion concentration when either an acid or a base is added to the solution. As an example, if only a few drops of concentrated hydrochloric acid are added to a beaker of pure water, the pH of the water might immediately fall to as low as 1.0. However, if a satisfactory buffer system is present, the hydrochloric acid combines instantaneously with the buffer, and the pH falls only slightly. Perhaps the best way to explain the action of an acid-base buffer is to consider an actual simple buffer system, such as the bicarbonate buffer, which is extremely important in regulation of acid-base balance in the body.

THE BICARBONATE BUFFER SYSTEM

A typical bicarbonate buffer system consists of a mixture of carbonic acid (H_2CO_3) and sodium bicarbonate ($NaHCO_3$) in the same solution. It must first be noted that carbonic acid is a very weak acid for two reasons: First, its degree of dissociation into hydrogen ions and bicarbonate ions is poor in comparison with that of many other acids. Second, about 399 parts of 400 of any carbonic acid in a solution almost immediately dissociate into carbon dioxide and water, the net result being a high concentration of dissolved carbon dioxide but only a low concentration of acid.

When a strong acid, such as hydrochloric acid, is added to a buffer solution containing bicarbonate salt, the following reaction takes place:

$$HCl + NaHCO_3 \rightarrow H_2CO_3 + NaCl \quad (2)$$

From this equation it can be seen that the strong hydrochloric acid is converted into the very weak carbonic acid. Therefore, addition of the HCl lowers the pH of the solution only slightly.

On the other hand, if a strong base, such as sodium hydroxide, is added to a buffer solution containing carbonic acid, the following reaction takes place:

$$NaOH + H_2CO_3 \rightarrow NaHCO_3 + H_2O \quad (3)$$

This equation shows that the hydroxyl ion of the sodium hydroxide combines with the hydrogen ion from the carbonic acid to form water and that the other product formed is sodium b

net result is exchange of the strong base NaOH for the weak base $NaHCO_3$.

Though this bicarbonate buffer system has been illustrated in the above reactions as a *mixture* of carbonic acid and *sodium* bicarbonate, in the intracellular fluid, where little *sodium* bicarbonate is present, the bicarbonate ion is provided mainly as *potassium* and *magnesium* bicarbonate.

Quantitative Dynamics of Buffer Systems

Dissociation of Carbonic Acid. All acids are ionized to a certain extent, and the percentage of ionization is called the *degree of dissociation*. Equation 4 illustrates the reversible relationship between undissociated carbonic acid and the two ions that it forms, H^+ and HCO_3^-

$$H_2CO_3 \rightleftharpoons H^+ + HCO_3^- \qquad (4)$$

A physicochemical law has been found to apply to the dissociation of all acids but as applied in this case specifically to carbonic acid, is expressed by the following formula.

$$\frac{H^+ \times HCO_3^-}{H_2CO_3} = K \qquad (5)$$

This formula states that in any given carbonic acid solution the concentration of hydrogen ions times the concentration of bicarbonate ions divided by the concentration of undissociated carbonic acid is equal to a constant, K.

However, it is almost impossible to measure the concentration of undissociated carbonic acid in a solution because it is also continually in reversible equilibrium with dissolved carbon dioxide in the solution. Ordinarily, the amount of dissolved carbon dioxide is approximately 400 times the concentration of the undissociated acid. On the other hand, it is possible to measure the total amount of dissolved carbon dioxide, and, since the amount of undissociated carbonic acid is proportional to the amount of dissolved carbon dioxide. Formula 5 above can also be expressed as follows:

$$\frac{H^+ \times HCO_3^-}{CO_2} = K' \qquad (6)$$

The only real difference between the preceding two formulas is that the constant K is approximately 400 times the constant K'.

Formula 6 can be changed into the following form:

$$H^+ = K' \cdot \frac{CO_2}{HCO_3^-} \qquad (7)$$

If we take the logarithm of each of the two sides of Formula 7, it becomes the following:

$$\log H^+ = \log K' + \log \frac{CO_2}{HCO_3^-} \qquad (8)$$

Now, the signs of the log H^+ and of the log K are changed from positive to negative and the carbon dioxide and bicarbonate are inverted in the last term, which is the same as changing its sign also, giving the following formula.

$$-\log H^+ = -\log K' + \log \frac{HCO_3^-}{CO_2}$$

It will be recalled from earlier in the chapter that $-\log H^+$ is equal to the pH of the solution. Likewise, $-\log K'$ *is called the pK of a buffer.* Therefore, this formula can be changed still further to the following:

$$pH = pK' + \log \frac{HCO_3^-}{CO_2} \qquad (9)$$

The Henderson-Hasselbalch Equation. For the bicarbonate buffer system the pK' is 6.1, and Formula 9 may be expressed as follows:

$$pH = 6.1 + \log \frac{HCO_3^-}{CO_2} \qquad (10)$$

This is called the Henderson-Hasselbalch equation, and by using it one can calculate the pH of a solution with reasonable accuracy if the molar concentrations of the carbonate ion and dissolved carbon dioxide are known. If the bicarbonate concentration is equal to the dissolved carbon dioxide concentration, the second member of the right-hand portion of the equation becomes log of 1, which is equal to zero. Therefore, for equal concentrations the pH of the solution is equal to the pK'.

From the Henderson-Hasselbalch equation one can readily see that *an increase in bicarbonate ion concentration causes the pH to rise*, or, in other words, shifts the acid-base balance toward the alkaline side. On the other hand, *an increase in the concentration of dissolved carbon dioxide decreases the pH*, or shifts the acid base balance toward the acid side. It will be apparent later in this chapter that one can change the concentration of dissolved CO_2 in the body fluids by increasing or decreasing the rate of respiration. In this way the respiratory system can change to a certain extent the pH of the body fluids. On the other hand, the kidneys can increase or decrease the concentration of bicarbonate ion in the body fluids, in this way increasing or decreasing the pH. Thus, these two major mechanisms for regulating hydrogen ion concentration operate principally by altering one or the other of the two elements of the bicarbonate buffer system.

The "Titration Curve" of the Bicarbonate Buffer System. Figure 37–1 shows the changes in pH of the body fluids when the ratio of bicarbonate ion to carbon dioxide changes. Note that when the concentrations of the two elements of the buffer are equal, the pH of the solution is 6.1, which is equal to the pK' of the bicarbonate buffer system. When base is added to the buffer a large proportion of the dissolved carbon dioxide and carbonic acid is converted into bicarbonate ions, and the ratio is altered. As a result, the pH rises as indicated by the forward slope of the curve. On the other hand, when acid is added, a large proportion of the bicarbonate ion is converted first into carbonic acid and then into dissolved carbon dioxide so that the pH falls as illustrated by the downslope of the curve.

Buffering Power of the Bicarbonate Buffer System. Referring once again to Figure 37–1, note that at the central point of the curve addition of a slight amount of acid or base causes minimal change in pH. However, toward each end of the curve addition of a slight amount of acid or base causes the pH to change greatly. Thus the so-called *buffering power* of the buffer system is

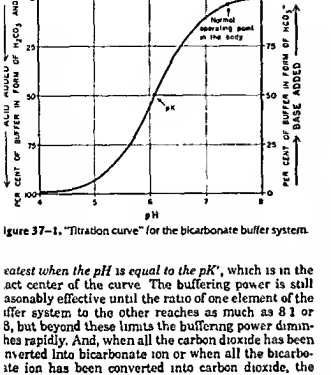


Figure 37–1. "Titration curve" for the bicarbonate buffer system.

eatest when the pH is equal to the pK', which is in the act center of the curve. The buffering power is still asonably effective until the ratio of one element of the uffer system to the other reaches as much as 8:1 or 8, but beyond these limits the buffering power diminhes rapidly. And, when all the carbon dioxide has been nverted into bicarbonate ion or when all the bicarbate ion has been converted into carbon dioxide, the stem has no more buffering power at all.

A second factor that determines the buffering power the concentrations of the two elements of the uffer olution, the CO_2 and HCO_3^-. Obviously, if the concentations are slight, only a small amount of acid or base dded to the solution will change the pH considerably. hus, *the buffering power of a buffer is also directly roportional to the concentrations of the buffer subances.*

HE BUFFER SYSTEMS F THE BODY FLUIDS

The three major buffer systems of the body fluids re the *bicarbonate buffer*, which was described bove, the *phosphate buffer*, and the *protein buffer*. ach of these performs major buffering functions nder different conditions.

The Bicarbonate Buffer System. The bicaronate system is not an exceedingly powerful uffer for two reasons. First, the pH in the extraellular fluids is about 7.4, while the pK' of the icarbonate buffer system is 6.1. This means that pproximately 20 times as much of the bicarbonate uffer is in the form of bicarbonate ion as is in the orm of dissolved carbon dioxide. For this reason, he system operates on a portion of its buffering urve where the buffering power is poor. Second, he concentrations of the two elements of the icarbonate system, CO_2 and HCO_3^-, are not great.

Yet, despite the fact that the bicarbonate buffer ystem is not especially powerful, it is actually more *mportant than all the others in the body because he concentration of each of the two elements of the icarbonate system can be regulated,* carbon dioxide oy the repiratory system and bicarbonate ion by the kidneys. As a result, the pH of the blood can be shifted up or down by the respiratory and renal regulatory systems.

The Phosphate Buffer System. The phosphate buffer system acts in almost identically the same manner as the bicarbonate buffer system, but it is composed of the following two elements: $H_2PO_4^-$ and HPO_4^{--}. When a strong acid, such as hydrochloric acid, is added to a mixture of these two substances, the following reaction occurs:

$$HCl + Na_2HPO_4 \rightarrow NaH_2PO_4 + NaCl \quad (12)$$

The net result of this reaction is that the hydrochloric acid is removed, and in its place an additional quantity of NaH_2PO_4 is formed. NaH_2PO_4 is only weakly acidic, so that the added strong acid is immediately traded for a very weak acid, and the pH changes relatively slightly.

Conversely, if a strong base, such as sodium hydroxide, is added to the buffer system, the following reaction takes place:

$$NaOH + NaH_2PO_4 \rightarrow Na_2HPO_4 + H_2O \quad (13)$$

Here sodium hydroxide is decomposed to form water and Na_2HPO_4. That is, a strong base is traded for the very weak base Na_2HPO_4, allowing only a slight shift in pH toward the alkaline side.

The phosphate buffer system has a pK of 6.8, which is not far from the normal pH of 7.4 in the body fluids; this allows the phosphate system to operate near its maximum buffering power. However, despite the fact that this buffer system operates in a reasonably good portion of the buffer curve, its concentration in the extracellular fluid is only one-twelfth that of the bicarbonate buffer. Therefore, its total buffering power in *the extracellular fluid* is even far less than that of the bicarbonate system.

On the other hand, the phosphate buffer is especially important in the tubular fluids of the kidneys for two reasons: First, phosphate usually becomes greatly concentrated in the tubules, thereby also greatly increasing the buffering power of the phosphate system. Second, the tubular fluid usually becomes more acidic than the extracellular fluid, bringing the operating range of the buffer closer to the pK of the system.

The phosphate buffer is also very important in the intracellular fluids because the concentration of phosphate in these fluids is many times that in the extracellular fluids, and also because the pH of the intracellular fluids is usually closer to the pK of the phosphate buffer system than is the pH of the extracellular fluid.

The Protein Buffer System. The most plentiful buffer of the body is the proteins of the cells and plasma, mainly because of their very high concentrations. There is a slight amount of diffusion of hydrogen ions through the cell membrane, and

even more important, carbon dioxide can diffuse within seconds through cell membranes and bicarbonate ions can diffuse to some extent (the hydrogen and bicarbonate ions require several hours to come to equilibrium in most cells other than the red blood cells). The diffusion of the two elements of the bicarbonate buffer system causes the pH in the intracellular fluids to change approximately in proportion to the changes in pH in the extracellular fluids. Thus, all the buffer systems inside the cells help buffer the extracellular fluids as well, but may take several hours to do so. These include the extremely large amounts of proteins inside the cells. Indeed, experimental studies have shown that about three quarters of all the *chemical* buffering power of the body fluids is inside the cells, and most of this results from the intracellular proteins. However, except for the red blood cells, the slowness of movement of hydrogen and bicarbonate ions through the cell membranes often delays the ability of the intracellular buffers to buffer extracellular acid-base abnormalities for several hours.

The method by which the protein buffer system operates is precisely the same as that of the bicarbonate buffer system. It will be recalled that a protein is composed of amino acids bound together by peptide linkages, but some of the different amino acids have free acidic radicals that can dissociate into base plus H^+. Furthermore, the pKs of a few of the protein buffering systems are not far from 7.4. This, too, helps make the protein buffering systems the most powerful of the body.

THE ISOHYDRIC PRINCIPLE

Each of the above buffer systems has been discussed as if it could operate individually in the body fluids. However, they all actually work together, for the hydrogen ion is common to the chemical reactions of all the systems. Therefore, whenever any condition causes the hydrogen ion concentration to change, it causes the balance of all the buffer systems to change at the same time. This phenomenon, called the *isohydric principle*, is presented by the following formula:

$$H^+ = \frac{K_1 \times HA_1}{A_1^-} = \frac{K_2 \times HA_2}{A_2^-} = \frac{K_3 \times HA_3}{A_3^-} \qquad (14)$$

in which K_1, K_2, and K_3 are the dissociation constants of three respective acids, HA_1, HA_2, and HA_3, and A_1^-, A_2^-, and A_3^- are the concentrations of the free negative ions that constitute the respective bases of the three buffer systems.

The important feature of this principle is that any condition that changes the balance of any one of the buffer systems also changes the balance of all the others, for *the buffer systems actually buffer each other by shifting hydrogen ions from one to the other.*

RESPIRATORY REGULATION OF ACID-BASE BALANCE

In the discussion of the Henderson-Hasselbalch equation, it was noted that an *increase in carbon dioxide concentration in the body fluids decreases the pH toward the acidic side*, whereas a *decrease in carbon dioxide raises the pH toward the alkaline side.* It is on the basis of this effect that the respiratory system is capable of altering the pH either up or down.

Balance Between Metabolic Formation of Carbon Dioxide and Pulmonary Expiration of Carbon Dioxide. Carbon dioxide is continually being formed in the body by the different intracellular metabolic processes, the carbon in the food being oxidized by oxygen to form carbon dioxide. This in turn diffuses into the interstitial fluids and blood and is transported to the lungs, where it diffuses into the alveoli and then is transferred to the atmosphere by pulmonary ventilation. However, several minutes are required for this passage of carbon dioxide from the cells to the atmosphere so that an average of 1.2 millimoles/liter of dissolved carbon dioxide is normally in the extracellular fluids at all times.

If the rate of metabolic formation of carbon dioxide becomes increased, its concentration in the extracellular fluids is likewise increased. Conversely, decreased metabolism decreases the carbon dioxide concentration.

On the other hand, if the rate of pulmonary ventilation is increased, carbon dioxide is blown off from the lungs and the amount of carbon dioxide in the extracellular fluids decreases.

Effect of Increasing or Decreasing Alveolar Ventilation on pH of Extracellular Fluids

If we assume that the rate of metabolic formation of carbon dioxide remains constant, then the only factor that affects the carbon dioxide concentration in the body fluids is the rate of the alveolar ventilation, as expressed by the following formula:

$$CO_2 \propto \frac{1}{\text{Alveolar ventilation}} \qquad (15)$$

And, since an increase in carbon dioxide decreases the pH, changes in alveolar ventilation also change the hydrogen ion concentration.

Figure 37–2 illustrates the approximate changes in pH in the blood that can be caused by increasing or decreasing the rate of alveolar ventilation. Note that an increase in alveolar ventilation to two times normal raises the pH of the extracellular fluids by about 0.23 pH unit. This means that if the pH of the body fluids had been 7.4 with normal

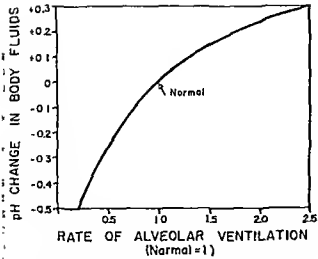


Figure 37–2. Approximate change in body fluid pH caused by or decreased rate of alveolar ventilation.

alveolar ventilation, doubling the ventilation raises the pH to 7.63. Conversely, a decrease in alveolar ventilation to one quarter normal reduces the pH 0.45 pH unit. That is, if at normal alveolar ventilation the pH had been 7.4, reducing the ventilation to one quarter reduces the pH to 6.95. Since alveolar ventilation can be reduced to zero ventilation or increased to about 15 times normal, one can readily understand how much the pH of the body fluids can be changed by alterations in the activity of the respiratory system.

Effect of Hydrogen Ion Concentration on Alveolar Ventilation

Not only does the rate of alveolar ventilation affect the hydrogen ion concentration of the body fluids, but, in turn, the hydrogen ion concentration affects the rate of alveolar ventilation. This results from a *direct action of hydrogen ions on the respiratory center in the medulla oblongata* that controls breathing, which will be discussed in detail in Chapter 42.

Figure 37–3 illustrates the changes in alveolar ventilation caused by changing the pH of arterial blood from 7.0 to 7.6. From this graph it is evident that a decrease in pH from the normal value of 7.4 to the strongly acidic level of 7.0 can increase the rate of alveolar ventilation to as much as four to five times normal, whereas an increase in pH into the alkaline range can decrease the rate of alveolar ventilation to only a fraction of the normal level.

Feedback Regulation of Hydrogen Ion Concentration by the Respiratory System. Because of the ability of the respiratory center to respond to hydrogen ion concentration, and because changes in alveolar ventilation in turn alter the hydrogen ion concentration in the body fluids, the respiratory system acts as a typical feedback regulatory system for controlling hydrogen ion concentration. That is, any time the hydrogen ion concentration becomes high, the respiratory system becomes more active, and alveolar ventilation increases. As a result, the carbon dioxide concentration in the extracellular fluids decreases, thus reducing the hydrogen concentration back toward normal. Conversely, if the hydrogen ion concentration falls too low, the respiratory center becomes depressed, alveolar ventilation also decreases, and the hydrogen ion concentration rises back toward normal.

Efficiency of Respiratory Regulation of Hydrogen Ion Concentration. Unfortunately, respiratory control cannot return the hydrogen ion concentration all the way to the normal value of 7.4 when some abnormality outside the respiratory system has altered the pH from the normal. The reason for this is that, as the pH returns toward normal, the stimulus that has been causing either increased or decreased respiration will itself begin to be lost. Ordinarily, the respiratory mechanism for regulation of hydrogen ion concentration has a

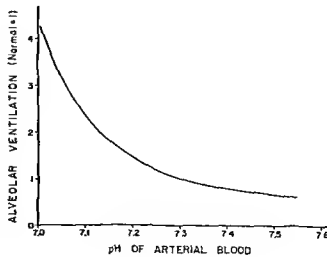


Figure 37–3. Effect of blood pH on the rate of alveolar ventilation. (Constructed from data obtained by Gray: Pulmonary Ventilation and Its Regulation. Charles C Thomas.)

control effectiveness of between 50 and 75 per cent (a feedback gain of 1 to 3). That is, if the hydrogen ion concentration should suddenly be decreased from 7.4 to 7.0 by some extraneous factor, the respiratory system, in 3 to 12 minutes, returns the pH to a value of about 7.2 to 7.3.

Buffering Power of the Respiratory System. In effect, respiratory regulation of acid-base balance is a *physiological type of buffer system* having almost identically the same importance as the chemical buffering systems of the body discussed earlier in the chapter. The overall "buffering power" of the respiratory system is one to two times as great as that of all the chemical buffers combined. That is, one to two times as much acid or base normally can be buffered by this mechanism as by the chemical buffers.

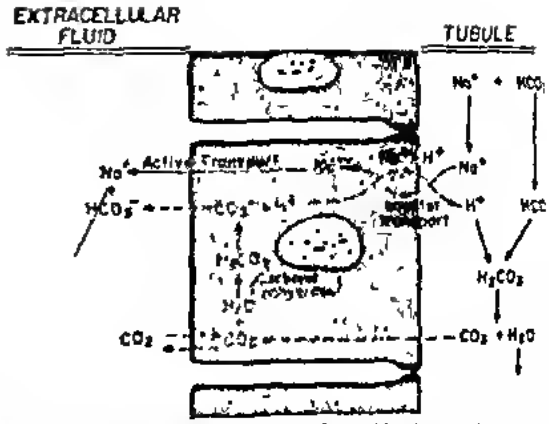


Figure 37–4. Chemical reactions for (1) hydrogen ion secre (2) sodium ion absorption in exchange for a hydrogen ion, (3) combination of hydrogen ions with bicarbonate ions in tubules

RENAL REGULATION OF HYDROGEN ION CONCENTRATION

In the earlier discussion of the Henderson-Hasselbalch equation it was pointed out that the kidneys regulate hydrogen ion concentration principally by increasing or decreasing the bicarbonate ion concentration in the body fluid. To do this, a complex series of reactions occurs in the renal tubules The following sections describe the different tubular mechanisms that accomplish these

TUBULAR SECRETION OF HYDROGEN IONS

The epithelial cells of the proximal tubules, thick segment of the ascending limb of the loop of Henle, distal tubules, collecting tubules, and collecting ducts all secrete hydrogen ions into the tubular fluid The mechanism by which this occurs is illustrated in Figure 37–4 The secretory process *begins with carbon dioxide* that either diffuses into or is formed by metabolism in the tubular epithelial cells. The carbon dioxide, under the influence of the enzyme *carbonic anhydrase* combines with water to form carbonic acid This then dissociates into *bicarbonate ion* and *hydrogen ion* Finally, the hydrogen ions are secreted by a mechanism of *Na^+-H^+ countertransport* That is, when sodium moves from the lumen of the tubule to the interior of the cell, it first combines with a *carrier protein* in the luminal border of the cell membrane, and at the same time a hydrogen ion inside the cell combines with the opposite side of the same carrier protein Then, because the concentration of sodium is much lower inside the cell than in the tubular lumen, this causes movement of the sodium down the concentration gradient to the interior, providing at the same time the energy for moving the hydrogen ion in the opposite direction ("counter" direction) into the tubular lumen. T process can even transport the hydrogen i against their concentration gradient.

In the collecting ducts hydrogen ion secret can continue until the concentration of hydro ions in the tubules becomes as much as 900 tu that in the extracellular fluid or, in other wo until the pH of the tubular fluids falls to ab 4.5 This represents a limit to the ability of tubular epithelium to secrete hydrogen ions.

About 84 per cent of all the hydrogen i secreted by the tubules are secreted in the pr mal tubules, but the maximum concentration g dient that can be achieved here is only about th to fourfold instead of the 900-fold that can achieved in the collecting ducts. That is, the can be decreased only to about 6 9, 0 5 pH u below 7.4, the pH of the glomerular filtrate. T distal tubules can decrease the pH to about 6 t 6 5, which is between that achieved by the pr mal tubules and collecting ducts.

Regulation of Hydrogen Ion Secretion the Carbon Dioxide Concentration in the I tracellular Fluid. Since the chemical reactu for secretion of hydrogen ions begin with carl dioxide, the greater the carbon dioxide concent tion, the more rapidly the reactions proceed, a the greater becomes the rate of hydrogen ion cretion. Therefore, any factor that increases carbon dioxide concentration in the extracellu fluids, such as decreased respiration or increa metabolic rate, also increases the rate of hydrog ion secretion. Conversely, any factor that decrea the carbon dioxide, such as excess pulmonary v tilation or decreased metabolic rate, decreases rate of hydrogen ion secretion.

At normal carbon dioxide concentrations, rate of hydrogen ion secretion is about 3 5 mi moles per minute, but this rises or falls directly proportion to changes in extracellular carbon di ide.

ıteraction of Bicarbonate Ions ıth Hydrogen Ions In the Tubules— Reabsorption" of Bicarbonate Ions

The renal tubules are not very permeable to the ıicarbonate ion because it is a large ion and also s electrically charged. However, the bicarbonate on can, in effect, be "reabsorbed" by a special ıocess which is also illustrated in Figure 37–4.

The reabsorption of bicarbonate ions is initiated ıy a reaction in the tubules between the bicarbo- ıate ions and the hydrogen ions secreted by the ubular cells, as illustrated in the figure. The ırbonic acid then dissociates into carbon dioxide ınd water. The carbon dioxide, having the capa- ıility of diffusing extremely readily through all ellular membranes, instantaneously diffuses hrough the epithelial cells all the way into the lood and is then carried to the lungs where it is ılown off into the air.

If we now noto in Figure 37–4 the chemical eactions that are responsible for formation of the ıydrogen ions in the epithelial cells, we see that ıach time a hydrogen ion is formed a bicarbonate on also is formed inside these cells by the disso- ıiation of H_2CO_3 into H^+ and HCO_3^-. This bicar- ɔonate ion then diffuses into the extracellular fluid n combination with a sodium ion that has been absorbed from the tubule.

Thus, the net effect of all these reactions is a mechanism for "reabsorption" of bicarbonate ions from the tubules, though the bicarbonate ions that enter the extracellular fluid are not the same bicarbonate ions that are removed from the tubular fluid.

Titration of Bicarbonate Ions Against Hydrogen Ions in the Tubules. Under normal conditions, the *rote of hydrogen ion secretion is about 3.50 millimoles/minute*, and the *rate of filtrotion of bicarbonate ions in the glomerular filtrote is about 3.49 millimoles/minute*. Thus, the quantities of the two ions entering the tubules are almost equal, and they combine with each other and actually annihilate each other, the end-products being carbon dioxide and water. Therefore, it is said that the bicarbonate ions and hydrogen ions normally "titrate" each other in the tubules.

However, note also that this titration process is not quite complete, for usually a slight excess of hydrogen ions (the acidic component) remains in the tubules to be excreted in the urine. The reason for this is that under normal conditions a person's metabolic processes continually form a small amount of excess acid that gives rise to the slight excess of hydrogen ions over bicarbonate ions in the tubules.

On rare occasions the bicarbonate ions are in excess, as we shall see in subsequent discussions. When this occurs, the titration process again is not quite complete; this time, excess bicarbonate ions (the basic component) are left in the tubules to pass into the urine.

Thus, the basic mechanism by which the kidney corrects either acidosis or alkalosis is by incomplete titration of hydrogen ions against bicarbonate ions, leaving one or the other of these to pass into the urine and therefore to be removed from the extracellular fluid.

From 85 to 98 per cent of this titration process occurs in the proximal tubules, and the carbonic acid formed by the titration reaction is then split very rapidly into its end-products, carbon dioxide and water. The water passes down the tubules, and the carbon dioxide diffuses into the extracellular fluid. To promote this rapid dissociation of carbonic acid into carbon dioxide and water, the luminal brush border surface of the proximal tubules (but not of the other tubules) has a large amount of attached carbonic anhydrase that accelerates the reaction.

RENAL CORRECTION OF ACIDOSIS— INCREASE OF BICARBONATE IONS IN THE EXTRACELLULAR FLUID

Now that we have described the mechanisms by which the renal tubules secrete hydrogen ions and reabsorb bicarbonate ions, we can explain the manner in which the kidneys readjust the pH of the extracellular fluids when it becomes abnormal.

The initial step in this explanation is to understand what happens to the concentrations of carbon dioxide and bicarbonate ions in the extracellular fluids in acidosis and alkalosis. First, let us consider *acidosis*. Referring again to Equation 11, the Henderson-Hasselbalch equation, we see that in *acidosis* the ratio of *carbon dioxide to bicarbonate ions* in the extracellular fluid increases. Therefore, the *rate of hydrogen ion secretion* rises to a level far greater than the *rate of bicarbonate ion filtration* into the tubules. As a result, a great excess of hydrogen ions is secreted into the tubules, which have far too few bicarbonate ions to react with. These excess hydrogen ions combine with the buffers in the tubular fluid, as explained in later paragraphs, and are excreted into the urine.

Figure 37–4 shows that each time a hydrogen ion is secreted into the tubules two other effects occur simultaneously: first, a bicarbonate ion is formed in the tubular epithelial cell and, second, a sodium ion is absorbed from the tubule into the epithelial cell. The sodium ion and bicarbonate ion then are transported together from the epithelial cell into the extracellular fluid.

Thus, *the net effect of secreting excess hydrogen ions into the tubules is to increase the quantity of bicarbonate ions in the extracellular fluid.* This increases the bicarbonate portion of the bicarbonate buffer system, which, in accordance with the Henderson-Hasselbalch equation and the isohydric principle, shifts all the buffers in the alkaline direction, increasing the pH in the process, and thereby correcting the acidosis.

RENAL CORRECTION OF ALKALOSIS—DECREASE OF BICARBONATE IONS IN THE EXTRACELLULAR FLUID

In *alkalosis*, the *ratio* of bicarbonate ions to dissolved carbon dioxide molecules increases. The effect of this on the titration process in the tubules is to increase the *ratio* of bicarbonate ions filtered into the tubules to hydrogen ions secreted. This increase occurs because the high extracellular bicarbonate ion concentration increases the bicarbonate ions filtered in the glomerular filtrate while at the same time the low carbon dioxide concentration in the extracellular fluid decreases the secretion of hydrogen ions. Therefore, the fine balance that normally exists in the tubules between the hydrogen and bicarbonate ions no longer occurs. Instead, far greater quantities of bicarbonate ions than hydrogen ions now enter the tubules. Since no bicarbonate ions can be reabsorbed without first reacting with hydrogen ions, all the excess bicarbonate ions pass into the urine and carry with them sodium ions or other positive ions. Thus, in effect, sodium bicarbonate is removed from the extracellular fluid

Loss of sodium bicarbonate from the extracellular fluid decreases the bicarbonate ion portion of the bicarbonate buffer system, and, in accordance with the Henderson-Hasselbalch equation, this shifts the pH of the body fluids back in the acid direction Furthermore, because of the isohydric principle, all the other body buffers shift back in the acid direction too Thus, the alkalosis is corrected.

COMBINATION OF EXCESS HYDROGEN IONS WITH TUBULAR BUFFERS AND TRANSPORT INTO THE URINE

When excess hydrogen ions are secreted into the tubules, only a small portion of these can be carried in the free form by the tubular fluid into the urine. The reason for this is that the maximum hydrogen ion concentration that the tubular system can achieve is $10^{-4.5}$ molar, which corresponds to a pH of 4.5. At normal daily urine flows, only 1 per cent of the daily excretion of excess hydrogen ions can be transported into the urine at this concentration.

Therefore, to carry the excess hydrogen ions into the urine the hydrogen ions must be carried in some other form than as free hydrogen ions. This is achieved by the hydrogen ions' first combining with introtubular buffers and then being transported in this form.

The tubular fluids have two very important buffer systems that transport the excess hydrogen ions into the urine: (1) the phosphate buffer and (2) the ammonia buffer. In addition, a number of weak buffer systems such as urate, citrate, and similar systems are of much less importance.

Transport of Excess Hydrogen Ions into the Urine by the Phosphate Buffer. The phosphate buffer is composed of a mixture of HPO_4^{--} and $H_2PO_4^-$. Both of these usually become considerably concentrated in the tubular fluid because of their relatively poor reabsorption and because of removal of water from the tubular fluid. Therefore even though the phosphate buffer is very weak in the blood, it is a much more powerful buffer in the tubular fluid.

Another factor that increases the importance of the phosphate buffer in the tubular fluids during acidosis is that the pK of this buffer is 6.8. When excess hydrogen ions are secreted, the tubular fluid normally begins at a pH of nearly 7.4 in the early proximal tubules and falls to about 6.0 in the distal tubules and collecting ducts Therefore, in these tubules the phosphate buffer functions in its most effective range, very near its pK value, as explained earlier in the chapter.

Figure 37–5 illustrates the manner in which hydrogen ions are removed from the tubular fluid by the phosphate buffer system and also how this functions in the total picture of renal acid-base control.

Transport of Excess Hydrogen Ions into the Urine by the Ammonia Buffer System. Another even more important, more complex, hydrogen ion buffer system of the tubular fluid is composed of ammonia (NH_3) and the ammonium ion (NH_4^+). The epithelial cells of all the tubules besides those of the thin segment of the loop of Henle continually synthesize ammonia, and this diffuses into the tubules. The ammonia then reacts with hydrogen ions, as illustrated in Figure 37–6, to form ammonium ions These are then excreted into the urine in combination with chloride ions and other tubular anions. Note in the figure that the net effect of these reactions is, again, *to increase the bicarbonate concentration* in the extracellular fluid

This ammonium ion mechanism for transport of excess hydrogen ions in the tubules is especially important for two reasons: (1) Each time an ammonia molecule combines with a hydrogen ion to form an ammonium ion the concentration of am-

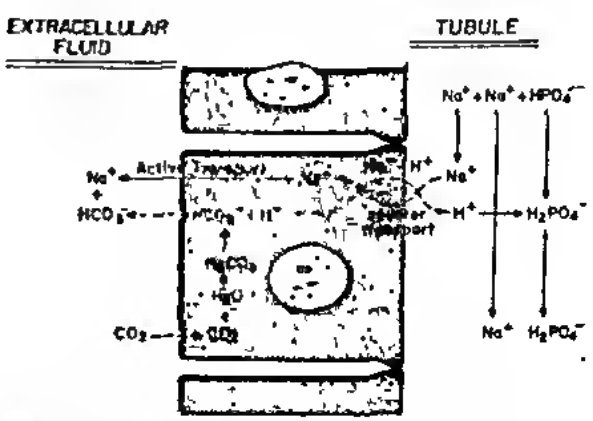


Figure 37–5. Chemical reactions in the tubules involving hydrogen ions, sodium ions, and the phosphate buffer system.

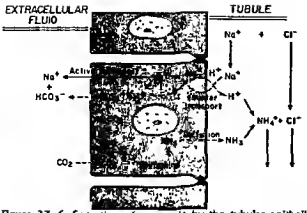


Figure 37–6. Secretion of ammonia by the tubular epithelial cells, and reaction of the ammonia with hydrogen ions in the tubules.

monia in the tubular fluid becomes decreased, which causes still more ammonia to diffuse from the epithelial cells into the tubular fluid. Thus, the rate of ammonia secretion into the tubular fluid is actually controlled by the amount of excess hydrogen ions to be transported. (2) Most of the negative ions of the tubular fluid are chloride ions. Only a few hydrogen ions could be transported into the urine in direct combination with chloride, because hydrochloric acid is a very strong acid and the tubular pH would fall rapidly below the *critical value of 4.5 below which further hydrogen ion secretion ceases.* However, when hydrogen ions combine with ammonia and the resulting ammonium ions then combine with chloride, the pH does not fall significantly because ammonium chloride is only very weakly acidic.

Sixty per cent of the ammonia secreted by the tubular epithelium is derived from *glutamine,* and the remaining 40 per cent from other amino acids or amines.

Enhancement of the Ammonia Buffer System in Chronic Acidosis. If the tubular fluids remain highly acidic for long periods of time, the formation of ammonia steadily increases during the first two to three days, rising as much as tenfold. For instance, immediately after acidosis begins, as little as 30 millimoles of ammonia might be secreted each day, but after several days as much as 300 to 450 millimoles can be secreted, illustrating that the ammonia-secreting mechanism can adapt readily to handle greatly increased loads of acid elimination. The principal cause of the increasing ammonia formation is that local acidosis of the tubular cells induces the production of large amounts of the enzyme *glutaminase* that is responsible for releasing ammonia from glutamine.

RAPIDITY OF ACID-BASE REGULATION BY THE KIDNEYS

The kidney mechanism for regulating acid-base balance cannot readjust the pH within seconds as can the extracellular fluid buffer systems nor within minutes as can the respiratory compensatory mechanism, but it is different from these other two in that it can keep on functioning for hours or days until it brings the pH almost exactly back to normal. In other words, its ultimate capability to regulate the body fluid pH, though slow to act, is infinitely more powerful than that of the other two regulatory mechanisms. The following few paragraphs will explain the quantitative importance of the kidneys for regulating the hydrogen ion concentration.

Figure 37–7 illustrates the effect of extracellular fluid pH on the rate at which bicarbonate ions are lost from or gained by the body fluids each minute. For instance, at a pH of 7.0, approximately 2.3 millimoles of bicarbonate ions are gained each minute, but as the pH returns toward the normal value of 7.4, the rate of gain falls to zero. Then, when the pH becomes significantly greater than 7.4, bicarbonate ions are lost by the extracellular fluids. For instance, at a pH of 7.6 about 1.5 millimoles of bicarbonate ions are lost each minute.

The total amount of buffers in the entire body (within the range of pH 7.0 to 7.8) is approximately 1000 millimoles. If all these should be suddenly shifted to the basic or acidic side by injecting a base or an acid, the kidneys would be able to return the pH of the body fluids back almost to normal in one to three days. But most important of all, this mechanism continues acting until the pH returns almost exactly to normal rather than a certain percentage of the way. Therefore, the real value of the renal mechanism for regulating hydrogen ion concentration is not rapidity of action but instead its ability in the end to neutralize completely any excess acid or alkali that enters the body fluids, unless the excess continues to enter.

Ordinarily, the kidneys can remove up to about 500 millimoles of either acid or base each day. If greater quantities than this enter the body fluids,

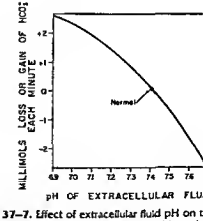


Figure 37–7. Effect of extracellular fluid pH on the rate of loss or gain of bicarbonate ions in the body fluids each minute

the kidneys are unable to cope with the extra load, and severe acidosis or alkalosis ensues.

Range of Urinary pH. In the process of adjusting the hydrogen ion concentration of the extracellular fluid, the kidneys often excrete urine at pHs as low as 4.5 or as high as 8.0. When acid is being excreted the urine pH falls, and when alkali is being excreted the pH rises. Even when the pH of the extracellular fluids is at the normal value of 7.4, a fraction of a millimole of acid is still lost each minute. The reason for this is that about 50 to 80 millimoles more acid than alkali are formed in the body each day, and this acid must be removed continually. Because of the presence of this excess acid in the urine, the normal urine pH is about 6.0 instead of 7.4, the pH of the blood.

RENAL REGULATION OF PLASMA CHLORIDE CONCENTRATION—THE CHLORIDE TO BICARBONATE RATIO

In the preceding discussions we have emphasized the ability of the kidneys to conserve bicarbonate ion in the extracellular fluids whenever a state of acidosis develops, or to remove bicarbonate ions in a state of alkalosis. Thus, the bicarbonate ion is shuttled back and forth between high and low values as one of the principal means of adjusting the acid-base balance of the extracellular buffer systems and therefore also for adjusting the extracellular fluid pH.

However, in the process of juggling the extracellular fluid concentration of bicarbonate ion, it is essential to remove some other anion from the extracellular fluids each time the bicarbonate is increased, or to increase some other anion when the bicarbonate concentration is decreased. In general, the anion that is reciprocally juggled up or down with the bicarbonate ion is the chloride ion because this is the anion in greatest concentration in the extracellular fluid.

CLINICAL ABNORMALITIES OF ACID-BASE BALANCE

RESPIRATORY ACIDOSIS AND ALKALOSIS

From the discussions earlier in the chapter it is obvious that any factor that decreases the rate of pulmonary ventilation increases the concentration of dissolved carbon dioxide in the extracellular fluid. And this in turn leads to increased carbonic acid and hydrogen ions, thus resulting in acidosis. Because this type of acidosis is caused by an abnormality of respiration, it is called *respiratory acidosis*.

On the other hand, excessive pulmonary ventilation reverses the process and decreases the hydrogen ion concentration, thus resulting in alkalosis; this condition is called *respiratory alkalosis*.

A person can cause respiratory acidosis in himself by simply holding his breath, which he can do until the pH of the body fluids falls to as low as perhaps 7.0. On the other hand, he can voluntarily overbreathe and cause alkalosis to a pH of about 7.9.

Respiratory acidosis frequently results from pathological conditions. For instance, damage to the respiratory center in the medulla oblongata that causes reduced breathing, obstruction of the passageways in the respiratory tract, pneumonia, decreased pulmonary membrane surface area, or any other factor that interferes with the exchange of gases between the blood and alveolar air can result in respiratory acidosis.

On the other hand, only rarely do pathologic conditions cause *respiratory alkalosis*. However, occasionally a psychoneurosis causes overbreathing to the extent that a person becomes alkalotic. Also, a physiologic type of respiratory alkalosis occurs when a person ascends to a *high altitude*. The low oxygen content of the air stimulates respiration, which causes excess loss of carbon dioxide and development of mild respiratory alkalosis.

METABOLIC ACIDOSIS AND ALKALOSIS

The terms *metabolic acidosis* and *metabolic alkalosis* refer to all other abnormalities of acid-base balance besides those caused by excess or insufficient carbon dioxide in the body fluids. Use of the word "metabolic" in this instance is unfortunate, because carbon dioxide is also a metabolic product. Yet, by convention, carbonic acid resulting from dissolved carbon dioxide is called a *respiratory acid* while any other acid in the body, whether it be formed by metabolism or simply ingested by the person, is called a *metabolic acid*, or a *fixed acid*.

Causes of Metabolic Acidosis. Metabolic acidosis can result from (1) failure of the kidneys to excrete the metabolic acids normally formed in the body, (2) formation of excessive quantities of metabolic acids in the body, (3) intravenous administration of metabolic acids, or (4) addition of metabolic acids by absorption from the gastrointestinal tract. Metabolic acidosis can result also from (5) loss of base from the body fluids. Some of the specific conditions that cause metabolic acidosis are the following:

Diarrhea. Severe diarrhea is one of the most frequent causes of metabolic acidosis for the following reasons: The gastrointestinal secretions normally contain large amounts of sodium bicarbonate. Therefore, excessive loss of these secretions during a bout of diarrhea is exactly the same as excretion of large amounts of sodium bicarbonate into the urine. In accordance with the Henderson-Hasselbalch equation, this results in a shift of the bicarbonate buffer system toward the acid side and results in metabolic acidosis. In fact, acidosis resulting from severe diarrhea can be so serious that it is one of the most common causes of death in young children.

Vomiting. A second cause of metabolic acidosis is vomiting. Vomiting of gastric contents alone, which sometimes occurs, obviously would cause loss of acid because the stomach secretions are highly acidic and would lead to alkalosis. But vomiting of contents from deeper in the gastrointestinal tract, which often occurs in much more quantity than loss of the stomach contents, will cause loss of alkali and result in metabolic acidosis.

Uremia. A third common type of acidosis is uremic acidosis, which occurs in severe renal disease. The cause of this is failure of the kidneys to rid the body of even the normal amounts of acids formed each day by the metabolic processes of the body.

Diabetes Mellitus. A fourth and extremely important cause of metabolic acidosis is diabetes mellitus. In this condition, lack of insulin secretion by the pancreas prevents normal use of glucose for metabolism. Instead, fat is split into acetoacetic acid, and this in turn is metabolized by the tissues for energy in place of glucose. Simultaneously, the concentration of acetoacetic acid in the extracellular fluids often rises very high, and large quantities of it are excreted in the urine, sometimes as much as 500 to 1000 millimoles per day.

Acidosis Caused by Carbonic Anhydrase Inhibitors. Administration of the common carbonic anhydrase inhibitor *acetazolamide* (Diamox), a drug that is frequently used to cause diuresis, also causes a mild degree of acidosis. This occurs because inhibition of the carbonic anhydrase on the luminal surface of the proximal tubular epithelium prevents adequate "reabsorption" of bicarbonate ions. Loss of these ions into the urine causes a decrease in bicarbonate ions in the extracellular fluid, thus leading to acidosis.

Causes of Metabolic Alkalosis. Metabolic alkalosis does not occur nearly as often as metabolic acidosis. However, there are several common causes of metabolic alkalosis, as follows:

Alkalosis Caused by Administering Diuretics (Except the Carbonic Anhydrase Inhibitors). All diuretics cause increased flow of fluid along the tubules, and this increase usually causes a great excess of sodium ions to flow into the distal and collecting tubules, leading also to rapid reabsorption of sodium ions from these tubules. This rapid reabsorption is coupled with enhanced hydrogen ion secretion because of the Na^+-H^+ countertransport mechanism in the luminal membranes of the tubular cells that links hydrogen secretion to sodium absorption; hence excessive loss of hydrogen ions from the body and resultant extracellular fluid alkalosis.

Excessive Ingestion of Alkaline Drugs. Perhaps the second most common cause of alkalosis is excessive ingestion of alkaline drugs, such as sodium bicarbonate, for the treatment of gastritis or peptic ulcer.

Alkalosis Caused by Loss of Chloride Ions. Excessive vomiting of gastric contents *without* vomiting of lower gastrointestinal contents causes excessive loss of hydrochloric acid secreted by the stomach mucosa. This loss leads to reduction of chloride ions and enhancement of bicarbonate ions in the extracellular fluid. The bicarbonate ions are derived from the stomach glandular cells that secrete the hydrogen ions, for these cells form bicarbonate ion in the same way that the tubular cells do when they secrete hydrogen ions. The net result is loss of acid from the extracellular fluids and development of metabolic alkalosis. This type of alkalosis occurs in newborn children who have pyloric obstruction caused by a greatly hypertrophied pyloric sphincter muscle.

Alkalosis Caused by Excess Aldosterone. When excess quantities of aldosterone are secreted by the adrenal glands, the extracellular fluid becomes slightly alkalotic. This is caused in the following way: The aldosterone promotes extensive reabsorption of sodium ions from the distal segments of the tubular system, but coupled with this is increased secretion of hydrogen ions (as explained earlier), which promotes alkalosis.

EFFECTS OF ACIDOSIS AND ALKALOSIS ON THE BODY

Acidosis. The major clinical effect of acidosis is depression of the *central nervous system*. When the pH of the blood falls below 7.0, the nervous system becomes so depressed that the person first becomes disoriented and, later, comatose. Therefore, patients dying of diabetic acidosis, uremic acidosis, and other types of acidosis usually die in a state of coma.

In metabolic acidosis the high hydrogen ion concentration causes increased rate and depth of respiration. Therefore, one of the diagnostic signs of metabolic acidosis is increased pulmonary ventilation. On the other hand, *in respiratory acidosis, respiration is usually depressed* because this is the cause of the acidosis, which is opposite to the effect in metabolic acidosis.

Alkalosis. The major clinical effect of alkalosis on the body is *overexcitability of the nervous system*. This occurs both in the central nervous system and in the peripheral nerves, but usually the peripheral nerves are affected before the central nervous system. The nerves sometimes become so excitable that they automatically and repetitively fire even when they are not stimulated by normal stimuli. As a result, the muscles go into a state of *tetany*, which means a state of tonic spasm. This tetany usually appears first in the muscles of the forearm, then spreads to the muscles of the face, and finally all over the body. Extremely alkalotic patients may die from tetany of the respiratory muscles.

Occasionally an alkalotic person develops severe symptoms of central nervous system overexcitability. The symptoms may manifest themselves as extreme nervousness or, in susceptible persons, as convulsions. For instance, in persons who are predisposed to epileptic seizures, simply overbreathing often results in an attack. Indeed, this is one of the clinical methods for assessing one's degree of epileptic predisposition.

RESPIRATORY COMPENSATION OF METABOLIC ACIDOSIS OR ALKALOSIS

It was pointed out earlier that the high hydrogen ion concentration of metabolic acidosis causes increased pulmonary ventilation, which in turn results in rapid removal of carbon dioxide from the body fluids and reduces the hydrogen ion concentration back toward normal. Thus, this respiratory effect helps to compensate for the metabolic acidosis. However, this compensation occurs only partway. Ordinarily, the respiratory system compensates between 50 and 75 per cent. That is, if the metabolic factor makes the pH of the blood fall to 7.0 at normal pulmonary ventilation, the rate of pulmonary ventilation normally increases sufficiently to return the pH of the blood to 7.2 to 7.3, as was pointed out earlier in the chapter.

The opposite effect occurs in metabolic acidosis. That is, alkalosis diminishes the pulmonary ventilation, which *in turn increases the hydrogen ion concentration* back toward normal. Here again the compensation can take place to about 50 to 75 per cent.

RENAL COMPENSATION OF RESPIRATORY ACIDOSIS OR ALKALOSIS

If a person develops respiratory acidosis that continues for a prolonged period of time, the kidneys will secrete an excess of hydrogen ions, resulting in an increase in sodium bicarbonate in the extracellular fluids. After one to six days, the pH of the body fluids will have returned about 65 to 75 per cent of the way to normal even though the person continues to breathe poorly.

Exactly the converse effect occurs in respiratory alkalosis. Large amounts of sodium bicarbonate are lost into the urine, decreasing the extracellular bicarbonate ion, thereby decreasing the pH also almost back to normal.

PHYSIOLOGY OF TREATMENT IN ACIDOSIS OR ALKALOSIS

Obviously, the best treatment for acidosis or alkalosis is to remove the condition causing the abnormality, but, if this cannot be effected, different drugs can be used to neutralize the excess acid or alkali.

To neutralize excess acid, large amounts of *sodium bicarbonate* can be ingested by mouth. This is absorbed into the blood stream and increases the bicarbonate ion portion of the bicarbonate buffer, thereby shifting the pH toward the alkaline side. Sodium bicarbonate is occasionally used also for intravenous therapy, but this has such strong and often dangerous physiological effects that other substances are often used instead, such as *sodium lactate* or *sodium gluconate*. The lactate and gluconate portions of the molecules are metabolized in the body, leaving the sodium in the extracellular fluids in the form of sodium bicarbonate, and thereby shifting the pH of the fluids in the alkaline direction.

For treatment of alkalosis, ammonium chloride is often administered by mouth. When this is absorbed into the blood, the ammonia portion of the ammonium chloride is converted by the liver into urea, this reaction liberates hydrochloric acid, which immediately reacts with the buffers of the body fluids to shift the hydrogen ion concentration in the acid direction. Occasionally, ammonium chloride is infused intravenously, but the ammonium ion is highly toxic and this procedure can be dangerous. Another substance occasionally used is *lysine monohydrochloride*.

CLINICAL MEASUREMENTS AND ANALYSIS OF ACID-BASE ABNORMALITIES

pH Measurements. In studying a patient with acidosis or alkalosis it is desirable to know the pH of the body fluids. This can be determined easily by measuring the pH of the plasma with a glass electrode pH meter. However, extreme care must be exercised in removing the plasma and in making the measurement, because even the slightest diffusion of carbon dioxide out of the plasma into the air shifts the bicarbonate buffer system in the alkaline direction, resulting in a much elevated pH measurement.

The pH-Bicarbonate Diagram. The so-called pH-bicarbonate diagram, which is illustrated in Figure 37–8, can be used to determine the *type* of acidosis or alkalosis a person has and its *severity*. Its use may be explained as follows.

The more vertical curves of the diagram depict different carbon dioxide concentrations. The normal carbon dioxide concentration of 1.2 mM/liter is denoted by the colored line (this is equivalent to a P_{CO_2} of 40 mm Hg). The points along this line represent the possible combinations of bicarbonate concentration and pH that can exist in the body fluids when the carbon dioxide concentration is normal.

The more horizontal lines represent the concentrations of metabolic acids and bases in the body fluids. The colored line, denoted by the numeral zero, indicates balance between both of these. That is, the points of

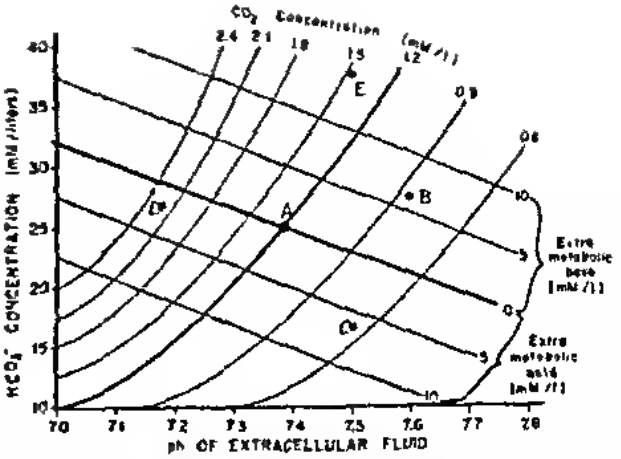


Figure 37–8. The pH-bicarbonate diagram for determining relative degrees of metabolic and respiratory acidosis or alkalosis in a patient. (Modified from Davenport: The ABC of Acid-Base Chemistry. University of Chicago Press.)

this line represent the possible combinations of bicarbonate concentration and pH that can occur as long as the metabolic acids and bases in the body fluids are normal. The upper two horizontal lines indicate respectively additions of 5 to 10 mM/liter of extra metabolic base to the body fluids, and the two lower horizontal lines indicate additions of 5 to 10 extra mM/liter of metabolic acid.

To use this diagram we simply find the pH of the blood and the bicarbonate concentration, then plot the appropriate point on the diagram. For instance, if the pH is found to be the normal value of 7.4 and bicarbonate concentration the normal value of 25 mM/liter, we plot point A, which represents the normal condition.

Now using data obtained from another patient, we plot a new point with a pH of 7.63 and a bicarbonate concentration of 28 mM/liter. This is point B on the diagram, which represents a carbon dioxide concentration of 0.8 mM/liter and 7 mM/liter extra metabolic base. Thus, this person has *metabolic alkalosis* because of the considerable extra metabolic base in the body fluids, but the person also has *respiratory alkalosis* because of overventilating to an extent that the carbon dioxide concentration is considerably less than normal.

Similarly, from other patients, we plot points C, D, and E. Point C represents 6 mM/liter of *metabolic acidosis* and sufficient *respiratory alkalosis* to reduce the carbon dioxide concentration to 0.7 mM/liter. A person with such a finding as this could have respiratory alkalosis that has been partially compensated by metabolic acidosis produced by the kidneys.

Point D represents *mild metabolic acidosis*, 2 mM/liter, combined with *severe respiratory acidosis*. A person could get into this state with severe primary respiratory acidosis and mild metabolic acidosis resulting from some other cause.

Point E represents *mild respiratory acidosis and severe metabolic alkalosis*. Presumably, in this case the metabolic alkalosis was primary, and respiratory compensation caused mild respiratory acidosis in an attempt to compensate for the metabolic alkalosis.

In summary, by using the pH-bicarbonate diagram, one can determine both the degree of metabolic acidosis or alkalosis and the degree of respiratory acidosis or alkalosis in a patient at the same time.

REFERENCES

Arruda, J A. L., and Kurtzman, N A. Relationship of renal sodium and water transport to hydrogen ion secretion *Annu Rev Physiol.*, *40* 43, 1978

Bank, N, *et al*. A micropuncture study of HCO_3 reabsorption by the hypertrophied proximal tubule. *Yale J. Biol Med*, *51*(3) 275, 1978

Brenner, B. M., and Stein, J. H (eds) Acid Base and Potassium Homeostasis. New York, Churchill Livingstone, 1978

Catto, G. R: Clinical Aspects of Renal Physiology Philadelphia, W B Saunders Company, 1981

Chan, J C: The influence of dietary intake on endogenous acid production. Theoretical and experimental background *Nutr Metab*, *16* 1, 1974

Christensen, H N · Body Fluids and the Acid Base Balance, 2nd Ed Philadelphia, W. B Saunders Co, 1964.

Goldberger, E A Primer of Water, Electrolyte, and Acid Base Syndromes. Philadelphia, Lea & Febiger, 1980

Gottschalk, C W., *et al*· Localization of urine acidification in the mammalian kidney, *Am J Physiol*, *198* 581, 1960

Guder, W G, and Schmidt. U (eds.) Biochemical Nephrology Bern, H Huber, 1978

Halperin, M. L., *et al.* Distal renal tubular acidosis syndromes A pathophysiological approach *Am J Nephrol*, *5* 1, 1985

Hills, A. G Acid Base Balance Baltimore, Williams & Wilkins, 1973

Irvine, R. O., Diet and drugs in renal acidosis and acid base regulation *Prog Biochem. Pharmacol*, *7* 146, 1972.

Ives, H. E., and Rector, F. C, Jr Proton transport and cell function *J Clin. Invest*, *73*.283, 1984

Jones, N. L. Blood Gases and Acid base Physiology New York, B C Decker, 1980.

Kintner, E. P; Acid base, blood gas, and electrolyte balances. *Prog Clin. Pathol*, *4* 143, 1972

Kovacevic, A., and McGivan, J D Mitochondrial metabolism of glutamine and glutamate and its physiological significance *Physiol Rev*, *63*.547, 1983.

Lemann, J, Jr, and Lennon, E J Role of diet, gastrointestinal tract, and bone in acid base homeostasis *Kidney Int*, *1* 275 1972

Lennon, E. J · Body buffering mechanisms *In* Frohlich, E D (ed) Pathophysiology, 2nd Ed. Philadelphia, J B Lippincott Co. 1976, p. 287

Levitin, H.: Acid-base balance *In* Bondy, P. K., and Rosenberg, L. E. (eds): Duncan's Diseases of Metabolism, 7th Ed Philadelphia, W B. Saunders Co, 1974, p 1531

Malmc, G.. Tubular handling of H *In* MTP International Review of Science: Physiology Vol 6 Baltimore, University Park Press, 1974, p. 107.

Malmc, G, and Giebisch, G · Symposium on acid base homeostasis. Mechanism of renal hydrogen ion secretion. *Kidney Int*, *1* 280, 1972

Pitts, R F Production and excretion of ammonia in relation to acid-base regulation. *In* Orloff, F, and Berliner, R. W (eds) Handbook of Physiology. Sec 8 Baltimore, Williams & Wilkins, 1973, p 455

Quintero, J A Acid Base Balance A Manual for Clinicians St. Louis, W. H. Green, 1979

Rahn, H Body temperature and acid base regulation (Review article) *Pneumonologie*, *151* 87, 1974

Rattenborg, C. C: Acid base monitoring *Clin Anesth*, *9* 139, 1973

Rector, F C., Jr Acidification of the urine *In* Orloff, F, and Berliner, R. W (eds.) Handbook of Physiology Sec 8 Baltimore, Williams & Wilkins, 1973, p 431

Relman, A S Metabolic consequences of acid-base disorders. *Kidney Int*, *1* 347, 1972.

Robinson, J R. Fundamentals of Acid Base Regulation, 4th Ed Philadelphia, J B Lippincott Co, 1972

Rose, B D Clinical Physiology of Acid-Base and Electrolyte Disorders. New York, McGraw-Hill Book Co, 1984

Siegg B K Symposium on acid base homeostasis The regulation of cerebrospinal fluid pH *Kidney Int*, *1* 360, 1972

Steinmetz, P R. Cellular mechanisms of urinary acidification *Physiol Rev*, *54*.890, 1974

Tannen, R. L. Control of acid secretion by the kidney *Annu Rev Med.*, *31* 35, 1980

Waddell, W J, and Bates, R. G Intracellular pH *Physiol Rev*, *49*:285, 1969

Weiner I M Transport of weak acid and bases *In* Orloff, F, and Berliner, R. W (eds) Handbook of Physiology Sec. 8 Baltimore, Williams & Wilkins, 1973, p 521

White, F N, and Somero, G Acid base regulation and phospholipid adaptations to temperature Time courses and physiological significance of modifying the milieu for protein function. *Physiol Rev*, *62* 40, 1982

38

Renal Disease, Diuresis, and Micturition

RENAL DISEASE

Renal disease can be classified into five different physiological categories: (1) *acute renal failure*, in which the kidneys stop working entirely or almost entirely, (2) *chronic renal failure*, in which progressively more nephrons are destroyed until the kidneys simply cannot perform all the necessary functions, (3) *hypertensive kidney disease*, in which vascular or glomerular lesions cause hypertension but not renal failure, (4) *nephrotic syndrome*, in which the glomeruli have become far more permeable than normal so that large amounts of protein are lost into the urine, and (5) *specific tubular abnormalities* that cause abnormal reabsorption or lack of reabsorption of certain substances by the tubules.

ACUTE RENAL FAILURE

Almost any condition that seriously interferes with kidney function can cause acute renal failure. Two of the most common causes are (1) acute glomerulonephritis and (2) acute damage and obstruction of the tubules.

Renal Failure Caused by Acute Glomerulonephritis. Acute glomerulonephritis is a disease caused by an abnormal immune reaction. In about 95 per cent of the patients, this occurs one to three weeks following an infection elsewhere in the body caused by certain types of Group A beta streptococci. The infection may have been a streptococcal sore throat, streptococcal tonsillitis, or even streptococcal infection of the skin. It is not the infection itself that causes damage to the kidneys. Instead, as antibodies develop during the succeeding few weeks against the streptococcal antigen, it is believed that the antibodies and antigen react with each other to form an insoluble immune complex that becomes entrapped in the glomerulus, especially in the basement membrane portion of the glomerulus.

Once the immune complex has deposited in the glomeruli, all the cells of the glomerulus begin to proliferate, but mainly the epithelial cells and the mesangial cells that lie between the endothelium and epithelium. In addition, large numbers of white blood cells become entrapped in the glomeruli. Many of the glomeruli become blocked entirely by this inflammatory reaction, and those that are not blocked usually become excessively permeable, allowing both protein and red blood cells to leak into the glomerular filtrate. In a few of the severest cases, either total or almost total renal shutdown occurs.

The acute inflammation of the glomeruli usually subsides in ten days to two weeks, and in most patients the kidneys return to normal function within the next few weeks to few months. Sometimes, however, many of the glomeruli are destroyed beyond repair, and in a small percentage of the patients progressive renal deterioration continues indefinitely, similar to that described for chronic glomerulonephritis in a subsequent section.

Tubular Necrosis as a Cause of Acute Renal Failure. Another common cause of acute renal shutdown is *tubular necrosis*, which means destruction of epithelial cells in the tubules, as illustrated in Figure 38–1. Common causes of tubular necrosis are (1) *various*

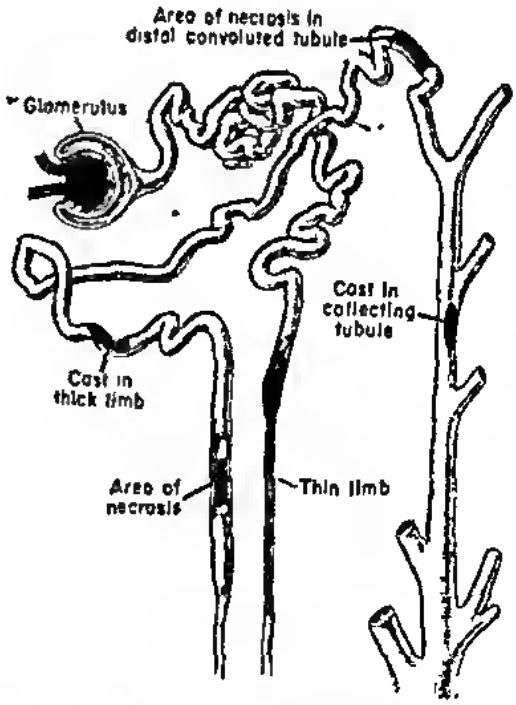


Figure 38–1. Damage to the distal tubules as a result of shock. (Modified from MacLean. Acute Renal Failure. Charles C Thomas.)

poisons that destroy the tubular epithelial cells, and (2) severe acute ischemia of the kidneys.

Renal Poisons. Among the different renal poisons are *carbon tetrachloride* and the heavy metals such as the *mercuric ion*. These substances have specific nephrotoxic action on the tubular epithelial cells, causing death of many of them. As a result, the epithelial cells slough away from the basement membrane and plug the tubules. In some instances the basement membrane also is destroyed, but, if not, new tubular epithelial cells can usually grow along the surface of the membrane so that the tubule becomes repaired within 10 to 20 days

Severe Acute Renal Ischemia. Severe ischemia of the kidney is likely to result from *severe circulatory shock*. In shock, the heart simply fails to pump sufficient amounts of blood to supply adequate nutrition to the different parts of the body, and renal blood flow is particularly likely to suffer because of strong sympathetic constriction of the renal vessels. Therefore, lack of adequate nutrition often destroys many tubular epithelial cells and thereby plugs many of the nephrons

Transfusion Reaction as a Cause of Acute Renal Failure. A transfusion reaction normally results in hemolysis of large amounts of red blood cells with release of hemoglobin into the plasma The size of the hemoglobin molecule is slightly less than that of the pores in the glomerular membrane so that much of the hemoglobin passes through the membrane into the glomerular filtrate. Therefore, after a transfusion reaction the tubular load of hemoglobin often becomes far greater than can be reabsorbed by the proximal tubules The excess hemoglobin is likely to precipitate in the nephron and in this way cause blockage. In addition, hemolysis of red blood cells probably also releases vasoconstrictor agents into the blood stream, and it is believed that the vasoconstriction can promote poor blood supply to the tubules, this acting as a further cause of tubular damage.

Physiological Effects of Acute Renal Failure

When the degree of acute renal failure is moderate, the major physiological effect is retention of salt and water. At first the tissues become edematous but the person has few other symptoms. However, within a few days the patient also develops hypertension, usually with a blood pressure increase of 30 to 40 mm Hg; this usually continues until the acute renal failure subsides.

In the severest cases, uremic retention of waste products ensues and acidosis soon develops. In complete renal shutdown without treatment, the patient dies within 8 to 14 days. Other effects of renal retention will be discussed in the following section in relation to chronic renal failure.

CHRONIC RENAL FAILURE—DECREASE IN NUMBER OF FUNCTIONAL NEPHRONS

In many types of kidney disease large numbers of nephrons are destroyed or damaged so severely that the remaining nephrons simply cannot perform the normal functions of the kidney. Some of the different causes of this include *chronic glomerulonephritis, traumatic loss of kidney tissue, congenital absence of kidney tissue, congenital polycystic disease* (in which large cysts develop in the kidneys and destroy surrounding nephrons by compression), *urinary tract obstruction* resulting from renal stones, *pyelonephritis*, and diseases of the renal vasculature.

Chronic Glomerulonephritis. Chronic glomerulonephritis is caused by any one of several different diseases that damage principally the glomeruli but often the tubules as well. The basic glomerular lesion is usually very similar to that which occurs in acute glomerulonephritis. It seems to begin with accumulation of precipitated antigen-antibody complex in the glomerular membrane, though in only a few instances is this caused by streptococcal infection. The result is inflammation of the glomeruli The glomerular membrane becomes progressively thickened and is eventually invaded by fibrous tissue. In the later stages of the disease, the glomerular filtration coefficient becomes greatly reduced because of decreased numbers of filtering capillaries in the glomerular tufts and because of thickened glomerular membranes. In the final stages of the disease many of the glomeruli are completely replaced by fibrous tissue, and the function of these nephrons is thereafter lost forever

Pyelonephritis. Pyelonephritis is an infectious and inflammatory process that usually begins in the renal pelvis but extends progressively into the renal parenchyma The infection can result from many different types of bacteria, but especially from the colon bacillus that originates from fecal contamination of the urinary tract. Invasion of the kidneys by these bacteria results in progressive destruction of renal tubules, glomeruli, and any other structures in the path of the invading organisms Consequently, large portions of the functional renal tissue are lost

A particularly interesting feature of pyelonephritis is that the invading infection usually affects the medulla of the kidney more than it affects the cortex. Since one of the primary functions of the medulla is to provide the counter-current mechanism for concentrating the urine, patients with pyelonephritis frequently have reasonably normal renal function except for reduced ability to concentrate their urine

Destruction of Nephrons by Renal Vascular Disease—Benign Nephrosclerosis. Though many different types of vascular lesions can lead to renal ischemia and death of the renal tissue, the most common of these are: (1) *atherosclerosis* of the larger renal arteries with progressive sclerotic constriction of the vessels, (2) *fibromuscular hyperplasia* of one or more of the large arteries, which also causes occlusion of these large vessels, and (3) *benign nephrosclerosis*, a very common condition caused by sclerotic lesions of the smaller arteries and arterioles.

Arteriosclerotic or hyperplastic lesions of the larger arteries frequently affect one kidney more than the other and therefore cause unilaterally diminished renal function.

Benign nephrosclerosis is believed to begin with leakage of plasma through the intimal membrane of the small arteries and arterioles. This causes a fibrinoid deposit to develop in the medial layers of these vessels followed by progressive fibrous tissue invasion that eventually constricts the vessel—in many instances totally occluding it. Since there is essentially no collateral circulation among the smaller renal arteries, destruction of one of these also causes destruction of a comparable amount of renal mass. Therefore, much of the kidney tissue becomes replaced by minute areas of fibrous tissue; the kidneys become greatly reduced in size and progressively develop a nodular surface. This process

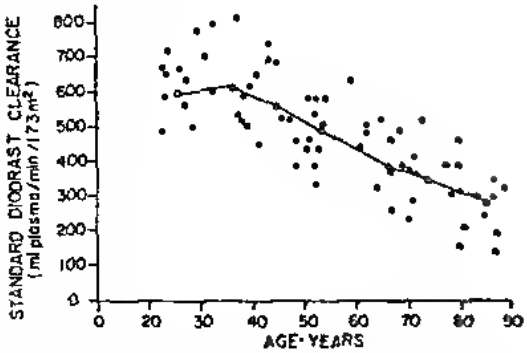


Figure 38–2. Effect of aging on the Diodrast clearance of the kidneys (Modified from Wolstenholme et al. Ciba Foundation Colloquia on Ageing Little, Brown and Co.)

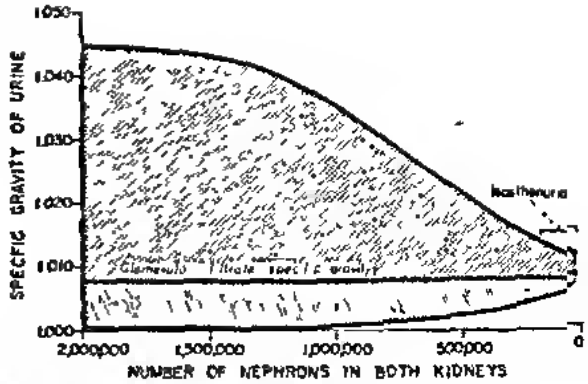


Figure 38–3. Development of *Isosthenuria* in patients with decreased numbers of active nephrons

occurs at least to some extent in almost all persons in old age and causes progressive decrease in both renal blood flow and renal plasma clearance. Figure 38–2 illustrates the Diodrast clearance of the kidneys (as a measure of renal blood flow) of otherwise normal persons at different ages, showing that even in the "normal" person, kidney blood flow decreases on the average to about 45 per cent of normal by the age of 80 years, and there is accompanying reduction of excretory function.

Abnormal Nephron Function in Chronic Renal Failure

Inability of Failing Kidneys to Maintain Normal Blood Constituents. The most important effect of renal failure is the inability of the kidneys to cope with large "loads" of electrolytes or other substances that must be excreted. Normally, only one third of the normal number of nephrons can eliminate essentially all the normal load of waste products from the body without serious accumulation of any of these in the body fluids. However, further reduction of numbers of nephrons leads to retention, especially of those waste products that depend on a high glomerular filtration rate for excretion. The most prominent of these are *urea* and *creatinine*. Death usually ensues when the number of nephrons falls below 5 to 20 per cent of normal.

Function of the Remaining Nephrons in Renal Failure. In renal failure, the still-functioning nephrons usually become greatly overloaded in several different ways. First, for reasons not understood, the blood flow through the glomerulus and the amount of glomerular filtrate formed by each nephron often increase 50 to 100 per cent. Second, extra large concentrations of excretory substances, such as urea, phosphates, sulfates, uric acid, and creatinine, accumulate in the extracellular fluid. On entering the glomerular filtrate of the few remaining functioning nephrons, these constitute markedly increased tubular loads of substances that are poorly reabsorbed. These increasing loads are partially compensated for by as much as 50 per cent increase in reabsorptive power of each tubule, but even this increase is often far too little to keep pace with the loads, which may be increased as much as 1000 per cent for each nephron because of the very high concentrations of the waste products in the plasma and glomerular filtrate. Therefore, only a small fraction of the solutes is reabsorbed, and the remaining solutes act like an osmotic diuretic, resulting in rapid flushing of tubular fluid [illegible] the tubules. Consequently, the volume of urine formation by each nephron can rise to as much as 20 times normal, and a person will occasionally have *as much as 2 to 3 times normal total urine output* despite significant renal insufficiency. This paradoxical situation is caused by a greater increase in urine volume output per nephron than reduction in numbers of nephrons.

Isosthenuria. Another effect of rapid flushing of fluid through the tubules is that the normal concentrating and diluting mechanisms of the kidney do not function properly. This occurs mainly because of too rapid flow of tubular fluid through the collecting ducts for adequate water absorption to occur. Therefore, as progressively more nephrons are destroyed, the specific gravity of the urine approaches that of the glomerular filtrate, approximately 1.008. These effects are illustrated in Figure 38–3, which gives the approximate upper and lower limits of urine specific gravity as the number of nephrons decreases. Since the concentrating mechanism is impaired more than the diluting mechanism, an important renal function test is to determine how well the kidneys can concentrate urine when the person is dehydrated for 12 or more hours.

Effects of Renal Failure on the Body Fluids—Uremia

The effect of renal failure on the body fluids depends to a great extent on the water and food intake of the person. Assuming that the person continues to ingest moderate amounts of water and food after complete renal shutdown, the concentration changes of different substances in the extracellular fluid are approximately those shown in Figure 38–4. The most important effects are (1) *generalized edema* resulting from water and salt retention, (2) *acidosis* resulting from failure of the kidneys to rid the body of normal acidic products, (3) *high concentrations of the nonprotein nitrogens*, especially *urea, creatinine*, and *uric acid*, resulting from failure of the body to excrete the metabolic end-products of proteins, and (4) *high concentration of other urinary retention products* including *phenols, guanidine bases, sulfates, phosphates*, and *potassium*. This condition is called *uremia* because of the high concentrations of normal urinary excretory products that collect in the body fluids.

Water Retention and Edema. If treatment is begun immediately after acute renal failure, the total body fluid content may not change at all. However, when the patient drinks water in response to normal desire, the body fluids begin increasing immediately and rapidly.

LIBRARY UDAIPUR

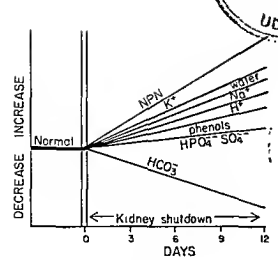


Figure 38–4. Effect of kidney shutdown on extracellular fluid constituents

If the patient takes in no electrolytes at the same time, approximately five eighths of the water enters the cells and three eighths remains in the extracellular fluids Thus, a patient with edema in renal failure will usually have both *intracellular* and *extracellular edema*. On the other hand, if the patient should ingest large amounts of salt along with the water, the edema might be only extracellular because of greatly increased amounts of osmotically active substances in the extracellular compartment.

Acidosis in Renal Failure. Each day the metabolic processes of the body normally produce 50 to 80 millimoles more metabolic acid than metabolic alkali Therefore, any time the kidneys fail to function, acid begins to accumulate In the body fluids Normally, the buffers of the fluids can buffer up to a total of 500 to 1000 millimoles of acid without severe depression of the extracellular fluid pH, and the phosphate compounds in the bones can buffer an additional few thousand millimoles, but gradually this buffering power is used up so that the pH falls drastically. The patient becomes *comatose* at about this same time, and it is believed that this is partly caused by the acidosis as is discussed subsequently.

Increase in Urea and Other Nonprotein Nitrogens (Azotemia) in Uremia. The nonprotein nitrogens include urea, uric acid, creatinine, and a few less important compounds. These, in general, are the end-products of protein metabolism and must be removed from the body continually to insure continued protein metabolism in the cells. The concentrations of these, particularly of urea, can rise to as high as ten times normal during one to two weeks of renal failure. However, even these high levels do not seem to affect physiological function nearly so much as do the high concentrations of hydrogen ions and some of the other less obvious substances such as very toxic guanidine bases, ammonium ions, and others Yet one of the most important means for assessing the degree of renal failure is to measure the concentrations of urea and creatinine.

Uremic Coma. After a week or more of renal failure the sensorium becomes clouded, and the patient soon progresses into a state of coma. The acidosis is believed to be one of the principal factors responsible for the coma because acidosis caused by other conditions, such as severe diabetes mellitus, also causes coma. However, many other abnormalities could also be contributory—the generalized edema, the high potassium concentration, and possibly even the high nonprotein nitrogen concentration.

The respiration usually is deep and rapid in coma, which is a respiratory attempt to compensate for the metabolic acidosis. In addition to this, during the last day or so before death, the arterial pressure falls progressively, then rapidly in the last few hours. Death occurs usually when the pH of the blood falls to about 6.8.

Anemia in Chronic Renal Failure. A patient with chronic renal failure of severe degree almost always also develops severe *anemia*. The probable cause of this is the following: The kidneys normally secrete the substance *erythropoietin* that in turn stimulates the bone marrow to produce red blood cells. Obviously, if the kidneys are seriously damaged they are unable to form adequate quantities of erythropoietin, which leads to diminished red blood cell production and consequent anemia However, other factors, such as the high plasma concentrations of urea, hydrogen ions, and other waste products might also play important roles in causing the anemia.

Osteomalacia in Renal Failure. Prolonged renal failure also causes *osteomalacia*, a condition in which the bones are partially absorbed and therefore greatly weakened, as will be explained in relation to the physiology of bone in Chapter 79 The most important cause of this condition is the following: Vitamin D must be converted by a two-stage process, first in the liver and then in the kidney, into 1,25-dihydroxycholecalciferol before it is able to promote calcium absorption from the intestine Therefore, serious damage to the kidneys greatly reduces the availability of calcium to the bones.

Dialysis of Patients with the Artificial Kidney

Artificial kidneys have now been used for almost 35 years to treat patients with severe renal failure. In certain types of acute renal failure, such as that following mercury poisoning or following circulatory shock, the artificial kidney is used simply to tide the patient over for a few weeks until the renal damage heals so that the kidneys can resume function. Also, the artificial kidney has now been developed to the point that many thousand persons with permanent renal failure or even total kidney removal are being maintained in health for years at a time, their lives depending entirely on the artificial kidney.

The basic principle of the *artificial kidney is to pass* blood through very minute blood channels bounded by a thin membrane. On the other side of the membrane is a *dialyzing fluid* into which unwanted substances in the blood pass by diffusion.

Figure 38–5 illustrates an artificial kidney in which blood flows continually between two thin sheets of cellophane; on the outside of the sheets is the dialyzing fluid. The cellophane is porous enough to allow all constituents of the plasma except the plasma proteins to diffuse freely in both directions—from plasma into the dialyzing fluid and from the dialyzing fluid back into the plasma. If the concentration of a substance is greater in the plasma than in the dialyzing fluid, there will be net transfer of the substance from the plasma into the dialyzing fluid. The amount of the substance that transfers depends on: (1) the difference between the

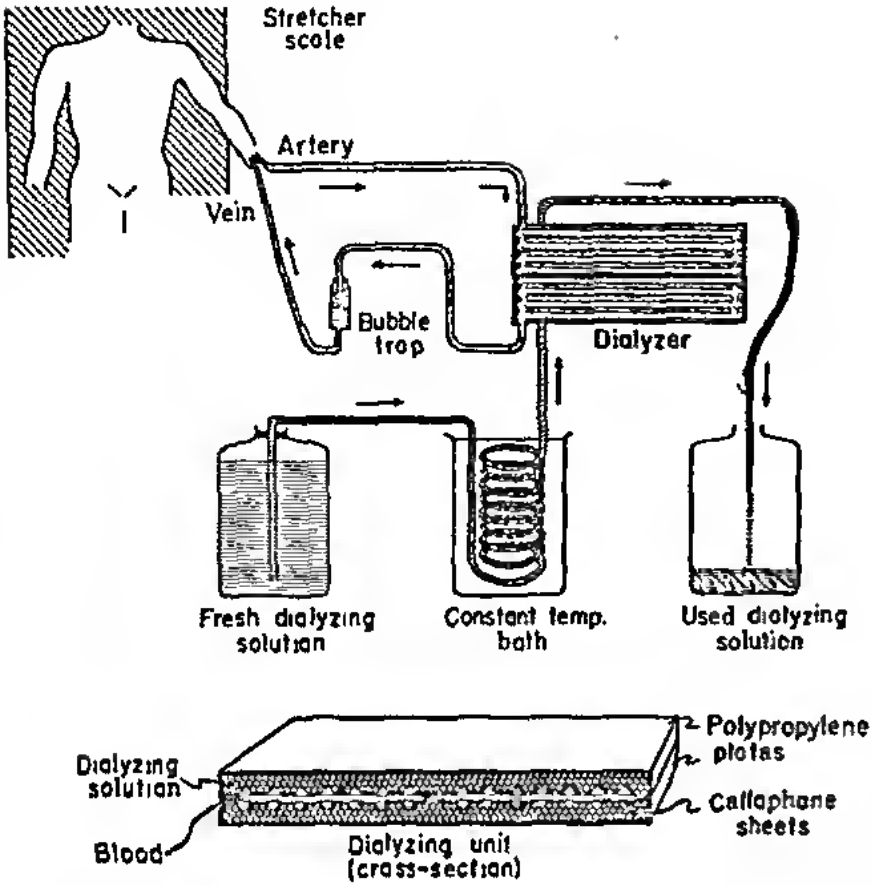


Figure 38–5. Principles of the artificial kidney

concentrations on the two sides of the membrane, (2) molecular size, the smaller molecules diffusing more rapidly than larger ones, and (3) the length of time that the blood and the fluid remain in contact with the membrane.

In normal operation of the artificial kidney, blood continually flows from an artery, through the kidney, and back into a vein. The total amount of blood in the artificial kidney at any one time is usually less than 500 milliliters, the rate of flow may be several hundred milliliters per minute, and the total diffusing surface is usually between 0.6 and 2.5 square meters. To prevent coagulation of blood in the artificial kidney, a small amount of heparin is infused into the blood as it enters the "kidney."

The Dialyzing Fluid. Table 38–1 compares the constituents in a typical dialyzing fluid with those in normal plasma and uremic plasma. Note that sodium, potassium, and chloride concentrations in the dialyzing fluid and in normal plasma are nearly identical, but in the uremic plasma the potassium concentration is considerably greater. This ion diffuses through the dialyzing membrane so rapidly that its concentration falls almost to equal that in the dialyzing fluid within only three to four hours' exposure to the dialyzing fluid.

On the other hand, there is no phosphate, urea, urate, sulfate, or creatinine in the dialyzing fluid, but they are present in high concentrations in the uremic blood. Therefore, when the uremic patient is dialyzed, these substances are lost in large quantities into the dialyzing fluid, thereby removing major proportions of them from the plasma.

Thus, the constituents of the dialyzing fluid are chosen that those substances in excess in the extracellular fluid in uremia can be removed at rapid rates, while the normal electrolytes remain essentially normal.

Effectiveness of the Artificial Kidney. The effectiveness of an artificial kidney is expressed in terms of the amount of plasma that can be cleared of different substances each minute, which, as will be recalled from Chapter 35, is also the principal means for expressing the functional effectiveness of the kidneys themselves. Most artificial kidneys can clear 100 to 225 milliliters of plasma per minute of urea, which shows that, at least in the excretion of this substance, the artificial kidney

Table 38–1. COMPARISON OF DIALYZING FLUID WITH NORMAL AND UREMIC PLASMA

Constituent	*Normal Plasma*	*Dialyzing Fluid*	*Uremic Plasma*
Electrolytes (mEq/liter)			
Na^+	142	142	142
K^+	5	4	7
Ca^{++}	3	3	2
Mg^{++}	1.5	1.5	1.5
Cl^-	107	107	107
HCO_3^-	27	27	14
Lactate$^-$	1.2	1.2	1.2
HPO_4^{--}	3	0	9
Urate$^-$	0.3	0	2
Sulfate^{--}	0.5	0	3
Nonelectrolytes (mg/dl)			
Glucose	100	125	100
Urea	26	0	200
Creatinine	1	0	6

can function about twice as rapidly as the two normal kidneys together, whose urea clearance is only 70 milliliters per minute. Yet the artificial kidney is normally used for only 4 to 6 hours three times a week. Therefore, the overall plasma clearance is still considerably limited when the artificial kidney replaces the normal kidneys.

HYPERTENSIVE KIDNEY DISEASE

Many of the same types of kidney disease that lead to chronic renal failure can also cause hypertension. However, this is not invariably true because damage to certain portions of the kidney is very prone to cause hypertension, whereas damage to other portions will cause uremia without hypertension. A classification of renal disease relative to its hypertensive or nonhypertensive effects is the following:

Hypertensive Renal Lesions. Essentially all renal lesions that cause *diminished blood flow* or *diminished glomerular filtration* per nephron will cause hypertension. Both these conditions tend to cause retention of salt and water, which eventually leads to hypertension (see Chapter 22). Once the hypertension has developed, the glomerular filtration rate may return entirely to normal. If all the tubules are normal, the filtrate is then processed normally in these so that urinary excretion may be completely normal, and there may be absolutely no sign of renal failure.

Renal Diseases That Lead to Renal Failure but Might Not Lead to Hypertension. Loss of large numbers of whole nephrons, such as occurs from loss of one kidney and part of another kidney, will always lead to renal failure if the amount of kidney tissue lost is great enough. However, if the remaining nephrons are completely normal, this condition frequently will not cause hypertension because even a slight rise in arterial pressure will increase the glomerular filtration rate sufficiently to promote rapid water and salt loss in the urine—even with few nephrons. On the other hand, a patient with this type of kidney abnormality who eats large quantities of salt will then develop very severe hypertension because the kidneys simply cannot clear adequate quantities of salt under these conditions.

Kidney Lesions That Tend to Cause Hypotension. An occasional person has a type of kidney disease that leads to tremendous salt and water loss in the urine even though there may be simultaneous retention of waste products and a state of uremia. This most frequently results from disease of the renal medulla, which prevents adequate absorption of fluid and electrolytes by the loops of Henle and collecting ducts. For instance, excessive long-term ingestion of aspirin and other analgesics can cause toxic destruction of the deeper portions of the medulla and lead to salt-wasting. As long as the person ingests sufficient quantities of salt, the pressure will remain normal. Placing a patient of this type on a low salt regime or on treatment with a diuretic, however, tends to decrease the arterial pressure. At the same time, excessive amounts of renin and angiotensin are usually formed, and this keeps the pressure from falling too much but leads to further uremic symptoms. In this state, any further stress to the arterial pressure control system can often lead to acute hypotension, culminating in death within a few hours.

Hypertension Caused by Renal Secretion of Renin. When one part of the kidney mass is ischemic while the remainder is nonischemic (such as occurs when one renal artery is severely constricted), the ischemic renal tissue secretes large quantities of renin. This secretion leads to the formation of angiotensin II, which in turn leads to hypertension. The most probable cause of the chronic hypertension, as was discussed in Chapter 22, is (1) the ischemic kidney tissue itself excretes less than normal amounts of water and salt, and (2) the angiotensin affects the nonischemic kidney tissue, causing it also to retain water and salt. (Many physiologists have believed in the past that peripheral vasoconstriction caused by the angiotensin was the cause of this type of hypertension. However, recent experiments have demonstrated almost conclusively that this is not the case, as was discussed in Chapter 22. In fact, the blood concentration of angiotensin is usually much too low to cause much vasoconstriction, but it can strongly affect renal excretion.)

THE NEPHROTIC SYNDROME—INCREASED GLOMERULAR PERMEABILITY

Large numbers of patients with renal disease develop a so-called *nephrotic syndrome*, which is characterized especially by *loss of large quantities of plasma proteins into the urine*. In some instances this occurs without evidence of any other abnormality of renal function, but more often it is associated with some degree of renal failure.

The cause of the protein loss in the urine is increased permeability of the glomerular membrane. Therefore, any disease condition that can increase the permeability of this membrane can cause the nephrotic syndrome. Such diseases include *chronic glomerulonephritis* (in the previous discussion, it was noted that this disease primarily affects the glomeruli and often causes a greatly increased permeability of the glomerular membrane), *amyloidosis*, which results from deposition of an abnormal proteinoid substance in the walls of the blood vessels and seriously damages the basement membrane of the glomerulus, and *lipoid nephrosis*, a disease found mainly in young children.

Lipoid Nephrosis. Lipoid nephrosis is also called "minimal change lesion" because one can rarely discern any abnormality of the glomerular membrane using the light microscope. It is very common in children, occurring most often before the age of 4 years but occurring occasionally in adults as well. Its basic cause is unknown, but the resulting renal lesion increases the permeability of the glomerular membrane and causes loss of proteins into the urine. This lesion, as seen in electron micrography, is the following: The epithelial cells that line the outer surface of the glomerulus have highly distorted foot processes. It will be recalled from the discussion of the glomerular membrane in Chapter 34 that the normal openings between these foot processes are very small slits, and it is believed that the smallness of these slits is a major factor in preventing the passage of protein through the glomerular membrane. Therefore, in the presence of abnormally shaped foot processes, tremendous quantities of protein leak into the tubules even though larger elements of the blood, such as red cells, usually are prevented from leaking.

The name "lipoid nephrosis" is derived from the fact that small quantities of lipid droplets are seen in the epithelial cells of the proximal tubules and also from the fact that the concentration of lipid substances in the

blood is usually increased (this is also true of some other types of nephrosis as well). The lipid deposits in the tubules apparently play no role in the increased loss of proteins.

The cause of lipoid nephrosis is unknown, but it frequently occurs soon after an upper respiratory infection caused by a virus or after other conditions that might elicit an immune response. Therefore, most research workers believe that some form of immune reaction is the cause of the basic glomerular lesion.

Administration of glucocorticoids such as prednisone will usually cause complete remission of the disease, although glucocorticoids will not cause remission of most other types of nephrotic syndrome.

Protein Loss. In the nephrotic syndrome, as much as 5 to 20 grams of plasma proteins can be lost into the urine each day. Though the resulting low plasma protein concentration stimulates the liver to produce far more plasma proteins than usual, nevertheless the liver often cannot keep up with the loss. Therefore, in severe nephrosis the colloid osmotic pressure sometimes falls extremely low, often from the normal level of 28 mm Hg to as low as 6 to 8 mm Hg.

Edema. The low colloid osmotic pressure in turn allows large amounts of fluid to filter into the interstitial spaces and also into the potential spaces of the body, thus causing serious *edema*. The untreated nephrotic person has been known on occasion to develop as much as 40 liters of excess extracellular fluid, and as much as 15 liters of this has been *ascites* in the abdomen. Also, the joints swell, and the pleural cavity and the pericardium become partially filled with fluid.

In severe nephrosis, the loss of fluid into the interstitial spaces causes the blood volume to diminish and the arterial pressure to fall 10 to 20 mm Hg below the pressure level in the absence of the nephrosis. These effects in turn lead to increased angiotensin and aldosterone production, both of which cause the kidneys to reabsorb excessive quantities of sodium chloride and water, which further compounds the seriousness of the edema.

A nephrotic person can be greatly benefited by intravenous infusion of large quantities of concentrated plasma proteins. Yet this is of only temporary benefit because enough protein can be lost into the urine in only a day or two to return the person to the original predicament.

SPECIFIC TUBULAR DISORDERS

In the discussion of active reabsorption and secretion by the tubules in Chapter 34, it was pointed out that the active transport processes are carried out by various carriers and enzymes in the tubular epithelial cells. Furthermore, there are a number of different carrier mechanisms for the different individual substances. In Chapter 3 it was also pointed out that each cellular enzyme and each carrier substance is formed in response to a respective gene or genes in the nucleus. If any required gene happens to be absent or abnormal, the tubules might be deficient in one of the appropriate enzymes or carriers. For this reason many different specific tubular disorders are known to occur for the transport of individual or special groups of substances through the tubular membrane. Essentially all of these are hereditary disorders. Some of the more important ones are.

Renal Glycosuria. In this condition the blood glucose concentration may be completely normal, but the transport maximum for reabsorption of glucose each minute is greatly limited. Consequently, despite the normal blood glucose level, large quantities of glucose pass into the urine each day. Because one of the tests for diabetes mellitus (which results from lack of insulin secretion by the pancreas) is the presence of glucose in the urine, renal glycosuria (a benign condition which causes essentially no dysfunction of the body) must always be ruled out before making a diagnosis of diabetes mellitus.

Nephrogenic Diabetes Insipidus. Occasionally the renal tubules do not respond well to antidiuretic hormone secreted by the supraoptirohypophysial system, and as a consequence large quantities of dilute urine are continually excreted. As long as the person is supplied with plenty of water, this condition rarely causes any severe difficulty. However, when adequate quantities of water are not ingested, the person rapidly becomes dehydrated.

Metabolic Acidosis Caused by Failure of the Tubules to Secrete Hydrogen Ions. In this condition the person is unable to secrete adequate quantities of hydrogen ions, and, as a result, large amounts of sodium bicarbonate are continually lost into the urine for reasons discussed in the preceding chapter. This causes a continual state of *metabolic acidosis*. However, adequate replacement therapy by continual administration of alkali can maintain normal bodily function.

Renal Hypophosphatemia. In renal hypophosphatemia the renal tubules fail to reabsorb adequate quantities of phosphate ions even when the phosphate concentration of the body fluids falls very low. This condition does not cause any serious immediate abnormalities, because the phosphate level of the extracellular fluids can vary widely without significant cellular dysfunction. However, over a long period of time the low phosphate level results in diminished calcification of the bones and causes the person to develop rickets. Furthermore, this type of rickets is refractory to vitamin D therapy in contrast to the rapid response of the usual type of rickets, which is discussed in Chapter 79.

Aminoacidurias. Some amino acids share mutual carrier systems for reabsorption, while other amino acids have distinct carrier systems of their own. Rarely, a condition called *generalized aminoaciduria* results from deficient reabsorption of all amino acids, but, more frequently, deficiencies of specific carrier systems may result in (1) *essential cystinuria*, in which large amounts of cystine fail to be reabsorbed and often crystallize in the urine to form renal stones, (2) simple *glycinuria*, in which glycine fails to be reabsorbed, or (3) *beta-aminoisobutyric aciduria*, which occurs in about 5 per cent of all people but apparently has no clinical significance.

RENAL FUNCTION TESTS

The renal function tests can be divided into three categories, (1) determination of renal clearances, (2) measurement of substances in the blood that are normally excreted by the kidneys, and (3) chemical and physical analyses of the urine.

Renal Clearance Tests. Any of the renal clearance tests, including clearance of para-aminohippuric acid, Diodrast, inulin, mannitol, or other substances, as described in Chapter 35, can be used as renal function

tests. Indeed, if all these are performed one can determine the glomerular filtration rate, the effective blood flow through the kidney per minute, the filtration fraction, and many other characteristics of renal function. However, because most of these clearance tests are difficult to do, several other clearance tests in which radiopaque or radioactive substances are excreted from the blood into the renal pelvis are usually employed. Some of these are the following:

Intravenous Pyelography. Several substances containing large quantities of iodine in their molecules—Diodrast, Hippuran, and Iopax—are excreted into the urine both by glomerular filtration and by active tubular secretion. Consequently, their concentration in the urine becomes very high within a few minutes after intravenous injection. Also, the iodine in the compounds makes them relatively opaque to x-rays. Therefore, x-ray pictures can be made showing shadows of the renal pelves, of the ureters, and even of the urinary bladder. Ordinarily, a sufficient quantity is excreted within five minutes after injection to give good shadows of the kidney pelves. Failure to show a distinct shadow within this time indicates depressed renal clearance.

Radioactive Clearance Studies. If any of the above substances (or many others) is prepared with *radioactive* iodine or some other radioactive nuclide, one can measure the radioactivity from both renal pelves by placing appropriate radioactivity counters over the kidneys. Only a trace amount of the substance need be injected intravenously and the degree of radioactivity recorded during the following few minutes to determine roughly the renal clearances.

A special value of both x-ray and radioactive pyelography is that they measure function of each kidney independently of the other rather than total function of both kidneys together, as is measured by the other renal function tests.

Blood Analyses as Tests of Renal Function. One can also estimate how well the kidneys are performing their functions by measuring the concentrations of various substances in the blood. For instance, the normal concentration of *urea* in the blood is 26 mg/dl, but in severe cases of renal insufficiency this can rise to as high as 300 mg/dl. The normal concentration of *creatinine* in the blood is 1.1 mg/dl, but this, too, can increase as much as tenfold. To determine the degree of metabolic acidosis resulting from renal dysfunction, one can use the *pH-bicarbonate diagram* or some other similar procedure, as discussed in the preceding chapter. Though these different tests are not as satisfactory as clearance tests for determining the functional capabilities of the kidneys, they are easy to perform, and they do tell the physician how seriously the internal environment has been disturbed.

Physical Measurements of the Urine as Renal Function Tests. Obviously, one of the most important urinary measurements is the *volume of urine* formed each day. In acute renal failure this can fall to zero, and in chronic renal failure it often is diminished. On the other hand, moderate renal failure may actually increase the urinary output, as was described earlier, because of vast overfunction of the remaining nephrons when the majority have been destroyed.

A second factor frequently measured is the *specific gravity* of the urine. Depending on the types of substances being cleared, this can vary tremendously; its upper extreme can go as high as 1.045 and it can fall to as little as 1.002. To test the ability of the kidneys to dilute the urine, the patient drinks large quantities of water, and measurements are made of the minimal specific gravity that can be attained. Then, at another time, the patient goes without water for 12 or more hours, and the maximum concentration of the urine is measured. Referring back to Figure 38–3, note that the concentrating ability of the kidneys is especially compromised as the number of nephrons decreases.

DIURETICS AND MECHANISMS OF THEIR ACTIONS

A diuretic is a substance that increases the rate of urine output. Most diuretics act by decreasing the rate of reabsorption of fluid from the tubules.

The principal use of diuretics is to reduce the total amount of fluid in the body. They are especially important in treating edema and hypertension.

When using a diuretic it is usually important that the rate of sodium loss in the urine also be increased as well as the rate of water loss. The reason for this is the following: If water alone were removed from the body fluids, the fluids would become hypertonic and elicit an osmoreceptor response, followed by marked secretion of antidiuretic hormone. This hormone would then cause large amounts of water to be reabsorbed from the tubules, which would nullify the effect of the diuretic. However, if sodium is lost along with the water, this nullification will not result. Therefore, all valuable diuretics caused marked *natriuresis* (sodium loss) as well as diuresis.

Several significant types of diuretics are the following:

The Osmotic Diuretics. Injection into the blood stream of *urea, sucrose, mannitol,* or any other substance not easily reabsorbed by the tubules causes a great increase in the osmotically active substances in the tubules. The osmotic pressure of these then greatly decreases water reabsorption, so that large amounts of tubular fluid flush on into the urine

The same effect occurs when the glucose concentration of the blood rises to very high levels in diabetes mellitus. Above a glucose concentration of about 250 mg/dl very little of the extra glucose is reabsorbed by the tubules; instead it acts as an osmotic diuretic and causes rapid loss of fluid into the urine. The name "diabetes" refers to the prolific urine flow.

Diuretics That Diminish Active Reabsorption. Any substance that inhibits carrier systems in the tubular epithelial cells and thereby diminishes active reabsorption of tubular solutes increases the tubular osmotic pressure and causes osmotic diuresis. Some of the most commonly used drugs of this type are the following:

The "Loop" Diuretics, Furosemide and Ethacrynic Acid. Furosemide and ethacrynic acid are the most powerful of all the clinically used diuretics. They are called loop diuretics because they block reabsorption of sodium chloride ions from the ascending limb of the loop of Henle as well as from all the remainder of the diluting segment of the distal tubule. This blockage of reabsorption of sodium and chloride ions causes diuresis for two reasons: (1) It allows greatly increased quantities of solutes to be delivered into the distal portions of the nephrons, and these then act as osmotic agents to

prevent water reabsorption as well (2) Failure to absorb sodium and chloride ions from the loop of Henle into the medullary interstitium decreases the concentration of the medullary interstitial fluid Consequently, the concentrating ability of the kidney is greatly decreased so that the reabsorption of fluid is further decreased Because of these two effects as much as 20 to 30 per cent of the glomerular filtrate may be delivered into the urine, causing under acute conditions urine outputs as great as 25 times normal for a period of a few minutes.

Chlorothiazide. Chlorothiazide and other thiazide derivatives act primarily on the distal tubules to prevent active sodium reabsorption; under favorable conditions they cause as much as 8 per cent of the glomerular filtrate to pass into the urine.

Carbonic Anhydrase Inhibitors—Acetazolamide. Acetazolamide (Diamox) and other carbonic anhydrase inhibitors primarily block reabsorption of bicarbonate ions from the proximal tubules. They do this by inhibiting the carbonic anhydrase that lines the luminal border of the tubular epithelial cells and normally catalyzes the dissociation of carbonic acid into water and carbon dioxide. This prevents removal of bicarbonate ions from the tubular fluid, and these remain in the tubules to act as an osmotic diuretic However, use of this drug also causes some degree of acidosis because of excessive loss of bicarbonate ions from the body fluids.

Competitive Aldosterone Inhibitors—Spironolactone. Spironolactone and several other similar substances compete with aldosterone for receptor sites in the distal nephron epithelial cells and thereby block the sodium reabsorption-promoting effect of aldosterone As a consequence, the sodium remains in the tubules and acts as an osmotic diuretic These drugs also block the effect of aldosterone to promote potassium secretion into the tubules. Therefore, in some instances, the extracellular fluid potassium concentration becomes dangerously elevated

MICTURITION

Micturition is the process by which the urinary bladder empties when it becomes filled. Basically the bladder (1) progressively fills until the tension in its walls rises above a threshold value, at which time (2) a nervous reflex called the "micturition reflex" occurs that either causes micturition or, if it fails in this, at least causes a conscious desire to urinate.

PHYSIOLOGIC ANATOMY AND NERVOUS CONNECTIONS OF THE BLADDER

The urinary bladder, illustrated in Figure 38–6, is a smooth muscle chamber composed of two principal parts (1) the *body*, which is the major part of the bladder, in which the urine collects, and (2) the *neck*, which is a funnel-shaped extension of the body, passing inferiorly and anteriorly into the urogenital triangle and connecting with the urethra. The bladder neck is also called the *posterior urethra* because of its relationship to the urethra.

The smooth muscle of the bladder is known as the *detrusor muscle*. Its muscle fibers extend in all directions when contracted, can increase the pressure in the bladder to as high as 40 to 60 mm Hg Thus,

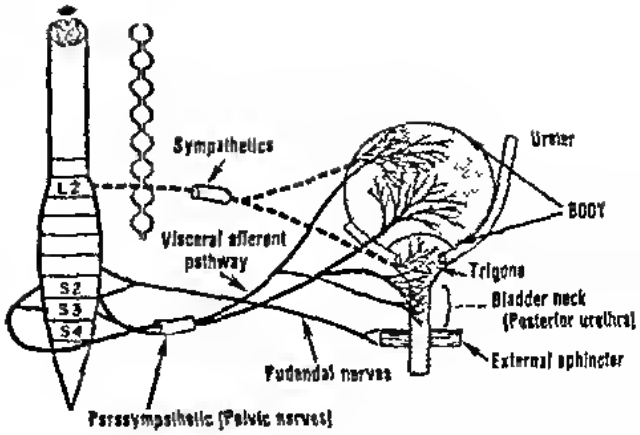


Figure 38–6. The urinary bladder and its innervation.

it is the detrusor muscle that empties the bladder. The smooth muscle cells of the detrusor muscle fuse with each other so that low resistance electrical pathways exist from one to the other. Therefore, an action potential can spread throughout the detrusor muscle to cause contraction of the entire bladder at once.

On the posterior wall of the bladder, lying immediately above the bladder neck, is a small triangular area called the *trigone* The lowermost apex of the trigone lies at the bladder neck, and the two ureters enter the bladder at the uppermost angles of the trigone. The trigone can be identified by the fact that its mucosa is very smooth, in contrast to the remainder of the bladder mucosa which is folded to form *rugae*. Where each ureter enters the bladder, it courses obliquely through the detrusor muscle and then passes still another 1 to 2 centimeters underneath the bladder mucosa before emptying into the bladder

The bladder neck (posterior urethra) is 2 to 3 centimeters long, and its wall is composed of detrusor muscle interlaced among a large amount of elastic tissue. The muscle in this area is frequently called the *internal sphincter*. Its natural tone normally keeps the bladder neck empty of urine, and therefore prevents emptying of the bladder until the pressure in the body of the bladder rises above a critical threshold.

Beyond the bladder neck, the urethra passes through the *urogenital diaphragm*, which contains a layer of muscle called the *external sphincter* of the bladder. This muscle is a voluntary skeletal muscle in contrast to the muscle of the bladder body and bladder neck, which is entirely smooth muscle. This external muscle is under voluntary control of the nervous system and can be used to prevent urination even when the involuntary controls are attempting to empty the bladder.

Innervation of the Bladder. The principal nerve supply to the bladder is by way of the *pelvic nerves*, which connect with the spinal cord through the sacral plexus, mainly connecting with cord segments S-2 and S-3 Coursing through the pelvic nerves are both *sensory nerve fibers* and *motor fibers*. The sensory fibers mainly detect the degree of stretch of the bladder walls Stretch signals from the bladder neck are especially strong and are mainly responsible for initiating the reflexes that cause bladder emptying.

The motor nerve fibers transmitted in the pelvic nerves are *parasympathetic fibers*. These terminate on ganglion cells located in the wall of the bladder. Short

postganglionic nerves then innervate the *detrusor muscle*.

Aside from the pelvic nerves, two other types of innervation are important to bladder function. Most important are the motor fibers transmitted through the *pudendal nerve* to the *external bladder sphincter*. These are somatic nerve fibers that innervate the skeletal muscle of this sphincter. In addition, the bladder receives sympathetic innervation from the sympathetic chain through the hypogastric nerves, connecting mainly with the L–2 segment of the spinal cord These sympathetic fibers probably stimulate mainly the blood vessels and have very little to do with bladder contraction. Some sensory nerve fibers also pass by way of the sympathetic nerves and may be important for the sensation of fullness and perhaps pain in some instances

TRANSPORT OF URINE THROUGH THE URETERS

The ureters are small smooth muscle tubes that originate in the pelves of the two kidneys and pass downward to enter the bladder. Each ureter is innervated by both sympathetic and parasympathetic nerves, and each also has an intramural plexus of neurons and nerve fibers that extends along its entire length

As urine collects in the pelvis, the pressure in the pelvis increases and initiates a peristaltic contraction beginning in the pelvis and spreading downward along the ureter to force urine toward the bladder. A peristaltic wave, traveling at a velocity of about 3 cm/sec, occurs from once every 10 seconds to once every 2 to 3 minutes The peristaltic wave can move urine against an obstruction with a pressure as high as 50 to 100 mm Hg. Transmission of the peristaltic wave is probably caused by action potentials passing along the smooth muscle syncytium of the ureteral wall. However, parasympathetic stimulation can increase and sympathetic stimulation can decrease the frequency of the waves and probably also can control the intensity of contraction

At the lower end, the ureter penetrates the bladder obliquely through the *trigone*, as illustrated in Figure 38–6. The ureter courses for several centimeters under the bladder epithelium so that pressure in the bladder compresses the ureter, thereby preventing backflow of urine when pressure builds up in the bladder during micturition.

Pain Sensations from the Ureters, and the Ureterorenal Reflex. The ureters are well supplied with pain nerve fibers. When the ureter becomes blocked, as by a ureteral stone, intense reflex constriction associated with very severe pain occurs. In addition, the pain impulses probably cause a sympathetic reflex back to the kidney to constrict the renal arterioles, thereby decreasing urinary output from that kidney. This effect is called the *ureterorenal reflex*; it obviously is important for prevention of excessive flow of fluid into the pelvis of a kidney with a blocked ureter.

TONE OF THE BLADDER WALL, AND THE CYSTOMETROGRAM DURING BLADDER FILLING

The solid curve of Figure 38–7 is called the *cystometrogram* of the bladder. It shows the changes in intravesical pressure as the bladder fills with urine. When no urine at all is in the bladder, the intravesical pressure is approximately zero, but by the time 30 to 50 milliliters of urine has collected, the pressure will have risen to 5 to 10 cm water. Additional urine up to 200 to 300 milliliters can collect with only a small amount of additional rise in pressure; this is caused by *intrinsic tone* of the bladder wall itself. Beyond 400 to 500 milliliters, collection of more urine causes the pressure to rise very rapidly.

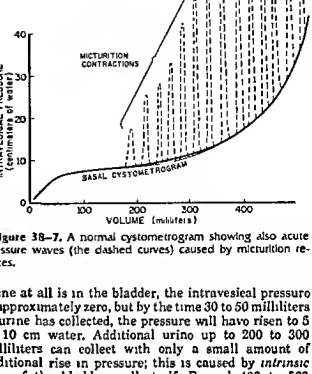


Figure 38–7. A normal cystometrogram showing also acute pressure waves (the dashed curves) caused by micturition reflexes.

Superimposed on the tonic pressure changes during filling of the bladder are periodic acute increases in pressure, which last from a few seconds to more than a minute The pressure can rise only a few centimeters of water or it can rise to over 100 cm water. These are *micturition waves* in the cystometrogram caused by the micturition reflex, which is discussed next.

THE MICTURITION REFLEX

Referring again to Figure 38–7, one sees that as the bladder fills, many superimposed *micturition contractions* begin to appear These are the result of a stretch reflex initiated by stretch receptors in the bladder wall, especially by the receptors in the bladder neck when this neck begins to fill with urine at the higher bladder pressures. *Sensory signals* are conducted to the sacral segments of the cord through the *pelvic nerves* and then back again to the bladder through the *parasympathetic fibers* in these same nerves.

Once a micturition reflex begins, it is "self-regenerative" That is, initial contraction of the bladder further activates the receptors to cause still further increase in afferent impulses from the bladder, which causes further increase in reflex contraction of the bladder, the cycle thus repeating itself again and again until the bladder has reached a strong degree of contraction. Then, after a few seconds to more than a minute, the reflex begins to fatigue, and the regenerative cycle of the micturition reflex ceases, allowing rapid reduction in bladder contraction. In other words, the micturition reflex is a single complete cycle of (a) progressive and rapid increase in pressure, (b) a period of sustained pressure, and (c) return of the pressure to the basal tonic pressure of the bladder. Once a micturition reflex has occurred and has not succeeded in emptying the bladder, the nervous

elements of this reflex usually remain in an inhibited state for at least a few minutes to sometimes as long as an hour or more before another micturition reflex occurs. However, as the bladder becomes more and more filled, micturition reflexes occur more and more often and more and more powerfully.

Once the micturition reflex becomes powerful enough and the fluid pressure in the bladder great enough to force the bladder neck open despite the tonic contraction of the bladder neck muscle, stretch of the neck not only greatly exacerbates the intensity of the micturition reflex itself but causes still another reflex as well This reflex passes to the sacral portion of the spinal cord and then back through the *pudendal nerve* to the *external sphincter* to inhibit it. If this inhibition is more potent than the voluntary constrictor signals from the brain, then urination will occur. If not, urination still will not occur until the bladder fills still more and the micturition reflex becomes more powerful.

Control of Micturition by the Brain. The micturition reflex is a completely automatic cord reflex, but it can be inhibited or facilitated by centers in the brain These include (a) strong *facilitatory and inhibitory centers in the brain stem*, probably located in the pons, and (b) several *centers located in the cerebral cortex* that are mainly inhibitory but can at times become excitatory.

The micturition reflex is the basic cause of micturition, but the higher centers normally exert final control of micturition by the following means:

1. The higher centers keep the micturition reflex partially inhibited all the time except when it is desired to micturate

2. The higher centers prevent micturition, even if a micturition reflex occurs, by continual tonic contraction of the external bladder sphincter until a convenient time presents itself.

3 When the time to urinate arrives, the cortical centers can (a) facilitate the sacral micturition centers to help initiate a micturition reflex, and (b) inhibit the external urinary sphincter so that urination can occur However, even more important, voluntary urination is usually initiated in the following way. First, the person contracts his abdominal muscles, which increases the pressure of the urine in the bladder At the same time, the pelvic floor muscles are relaxed, which lengthens the bladder neck and allows extra urine to enter the neck under pressure, thus stretching its walls This then excites the stretch receptors, which excites the micturition reflex and simultaneously inhibits the external urethral sphincter. Ordinarily, all the urine will then be emptied, with rarely more than 5 to 10 milliliters left in the bladder.

ABNORMALITIES OF MICTURITION

The Atonic Bladder. Destruction of the sensory nerve fibers from the bladder to the spinal cord prevents transmission of stretch signals from the bladder and, therefore, also prevents micturition reflex contractions Therefore, the person loses all bladder control despite intact efferent fibers from the cord to the bladder and despite intact neurogenic connections with the brain Instead of emptying periodically, the bladder fills to capacity and overflows a few drops at a time through the urethra. This is called *overflow dribbling*.

The atonic bladder was a common occurrence when syphilis was widespread, because syphilis frequently causes constrictive fibrosis around the dorsal nerve root fibers where they enter the spinal cord and, subsequently, destroys these fibers. This condition is called *tabes dorsalis*, and the resulting bladder condition is called a *tabetic bladder*. Another common cause of this condition is crushing injuries to the sacral region of the cord

The Automatic Bladder. If the spinal cord is damaged above the sacral region but the sacral segments are still intact, typical micturition reflexes still occur. However, they are no longer controllable by the brain. During the first few days to several weeks after the damage to the cord has occurred, the micturition reflexes are completely suppressed because of the state of "spinal shock" caused by the sudden loss of facilitatory impulses from the brain stem and cerebrum. However, if the bladder is emptied periodically by catheterization to prevent physical bladder injury, the excitability of the micturition reflex gradually increases until typical micturition reflexes return

It is especially interesting that stimulating the skin in the genital region can sometimes elicit a micturition reflex in this condition, thus providing a means by which some patients can still control urination.

The Uninhibited Neurogenic Bladder. Another common abnormality of micturition is the so-called "uninhibited neurogenic bladder," which results in frequent and relatively uncontrollable micturition This condition derives from partial damage in the spinal cord or brain stem that interrupts most of the inhibitory signals. Therefore, facilitatory impulses passing continually down the cord keep the sacral centers so excitable that even a very small quantity of urine will elicit an uncontrollable micturition reflex and thereby promote urination.

REFERENCES

Adler, S G., *et al* Hypersensitivity phenomena and the kidney: Role of drugs and environmental agents *Am. J Kidney Dis*, *5* 75, 1985

Avram, M M. (ed) Prevention of Kidney Disease and Long Term Survival New York, Plenum Publishing Corp, 1982

Bernstein, K N, and O'Connor, D T Antiadrenergic antihypertensive drugs. Their effect on renal function. *Annu. Rev Pharmacol. Toxicol*, *24* 105, 1984

Boyarsky, S *et al* Care of the Patient with Neurogenic Bladder Boston, Little, Brown, 1979

Brater, D C Pharmacodynamic considerations in the use of diuretics. *Annu. Rev Pharmacol Toxicol*, *23* 45, 1983

Bricker, N S. The Kidney Diagnosis and Management New York, John Wiley & Sons 1984

Brun, C., and Olsen, S Atlas of Renal Biopsy Philadelphia, W B. Saunders Company, 1981

Bullock, M L., *et al*: The assessment of risk factors in 462 patients with acute renal failure *Am. J Kidney Dis*, *5* 97, 1985

Burg, M, and Stoner, L. Renal tubular chloride transport and the mode of action of some diuretics *Annu Rev Physiol*, *38* 37, 1976

Catto, G E Clinical Aspects of Renal Physiology Philadelphia, W B Saunders Company, 1981

Chapman, A. (ed.) Acute Renal Failure New York, Churchill Livingstone, 1979

Charlton, C. A.. The Urological System New York, Churchill Livingstone, 1983

Churg, J: Renal Disease—Classification and Atlas of Glomerular Diseases. New York, Igaku-Shoin Medical Publishers, 1982

Coleman, T G, *et al*: Chronic hemodialysis and circulatory function *Simulation, Nov* 222, 1970

Conger, J D, and Schrier, R. W., Renal hemodynamics in acute renal failure *Annu Rev Physiol*, *42* 603, 1980

Cummings, N. B (ed.) Immune Mechanisms in Renal Disease New York, Plenum Publishing Corp, 1983

Darmady, E M, and MacIver, A G., Renal Pathology Boston, Butterworths, 1979

DePronzo, R. A., *et al*: Clinical disorders of hyperkalemia. *Annu. Rev. Med.*, 33:521, 1982
Diem, J., *et al.*: Studies on possible mechanisms of early functional compensatory adaptations in the remaining kidney. *Yale J. Biol Med*, 51(3) 265, 1978.
Dirks, J. H., and Wong, N. L. M: Acute functional adaptation to nephron loss Micropuncture studies. *Yale J. Biol. Med.*, 51(3) 255, 1978
Drukker, W., *et al.* (eds). Replacement of Renal Function by Dialysis. Boston, M. Nijhoff Medical Division, 1979.
Earley, L. E, and Gottschalk, C. W. (eds). Strauss and Welt's Diseases of the Kidney, 3rd Ed Boston, Little, Brown, 1979.
Cinaler, E. M., *et al*: The natural history and response to therapy of lupus nephritis. *Annu. Rev. Med*, 31 463, 1980
Gorevic, P. D, and Franklin, E. C: Amyloidosis. *Annu. Rev Med.*, 32.261, 1981.
Crantham, J. J: Fluid secretion in the nephron: Relation to renal failure. *Physiol. Rev*, 56.248, 1976
Harrington, J. T, and Kassirer, J. P.. Renal vein thrombosis *Annu. Rev Med*, 33 255, 1982.
Herrin, J T: Management of children with progressiva renal failure. *Annu. Rev Med.*, 34.21, 1983
Juncos, L. I: Physiological Basis of Diuretic Therapy: A Programmed Course. Springfield, Ill., Charles C Thomas, 1978
Kahan, B D., *et al.*. Clinical and experimantal studies with cyclosporine in renal transplantation. *Surgery*, 97 125, 1985
Kanwar, Y. S: Biophysiology of glomerular filtration and proteinuria. *Lab Invest*, 51:7, 1984.
Kincaid-Smith, P., *et al* (eds.) Glomerulonephritis. Morphology, Natural History, and Treatment. New York, John Wiley & Sons, 1973
Kincaid-Smith, P, *et al* (eds.) Progress in Glomerulonephritis. New York, John Wiley & Sons, 1979
Knox, F. G. (ed.): Textbook of Renal Pathophysiology Hagerstown, Md, Harper & Row, 1978
Krane, R. J, and Siroky, M B (eds.): Clinical Neuro-Urology. Boston, Little, Brown, 1979.
Leaf, A., and Cotran, R. S: Renal Pathophysiology 2nd Ed. New York, Oxford University Press, 1980.
Leaf, A., and Giebisch, G. (eds.) Renal Pathophysiology—Recent Advances. New York, Raven Press, 1979
Martinez, M. M: Handbook of Renal Therepeutics. New York, Plenum Publishing Corp., 1983.
Opelz, G, and Lenhard, V.: Immunological factore influencing renal graft survival. *Annu. Rev Med.*, 34 133, 1983.
Pavon, M M. (ed.) Cancer of the Prostate and Kidney. New York, Plenum Publishing Corp, 1983.
Pesce, A. J., and First, M. R.. Proteinuria: An Integrated Review New York, Marcel Dekker, 1979.
Porter, G. (ed.): Nephrotoxic Mechanisms of Drugs and Environmental Toxins New York, Plenum Publishing Corp., 1982.
Reidenberg, M M, and Drayer, D E.: Drug therepy in renal failure. *Annu. Rev Pharmacol. Toxicol*, 20.45, 1980.
Rindon, R. A.. Atlas of Renal Pathology. Philadelphia, J. B Lippincott Co, 1981
Simmah, R., and Cohen, A. H Glomerular capillary aneurysms in light-chain nephropathy An ultrastructural proposal of pathogenesis. *Am. J Pathol*, 118 293, 1985.
Strauss, J (ed) Nephrotic Syndrome New York, Garland Press, 1979.
Sutton, R. A.. Diuretics and calcium metabolism. *Am. J Kidney Dis.*, 5 4, 1985
Thurau, K. (ed.). Experimental Arute Renal Failure. New York, Springer-Verlag, 1976
Wagner, R. Nephrotic Syndrome Contemporary Patient Management. New Hyde Park, N Y, Medical Examination Publishing Co., 1981
Walser, M. Nutrition in renal failure. *Annu. Rev. Nutr*, 3.125, 1983
Wardle, E N Renal Medicine Baltimore, University Park Press, 1979
Warren, S, Jr (ed.) Self-Assessment of Current Knowledge in Nephrology New Hyde Park, N Y, Medical Examination Publishing Co., 1981
Weller, J. M (ed.) Fundamentals of Nephrology. Hagerstown, Md., Harper & Row, 1979
Williams, R. C, Jr Immune complexes in human diseases *Annu. Rev Med.*, 32.13, 1981
Wills, M R. Uremic toxins, and their effect on intermediary metabolism. *Clin Chem.*, 31 5, 1985
Wilson, C B (ed.) Immunological Mechanisms of Renal Disease New York, Churchill Livingstone, 1979
Wood, R. F: Renal Transplantation. A Clinical Handbook. Philadelphia, W B Saunders Company, 1984
Yoshitosha, Y (ed.). Glomerulonephritis. Baltimore, University Park Press, 1979.
Zmijewski, C. M.. Human leukocyte antigen matching in renal transplantation. Review and current status. *J Surg Res*, 38 66, 1985

VII

RESPIRATION

39 ■ Pulmonary Ventilation

40 ■ Physical Principles of Gaseous Exchange; Diffusion of Oxygen and Carbon Dioxide Through the Respiratory Membrane

41 ■ Transport of Oxygen and Carbon Dioxide in the Blood and Body Fluids

42 ■ Regulation of Respiration

43 ■ Respiratory Insufficiency—Pathophysiology, Diagnosis, Oxygen Therapy

39

Pulmonary Ventilation

The process of respiration can be divided into four major mechanistic events. (1) pulmonary ventilation, which means the inflow and outflow of air between the atmosphere and the lung alveoli, (2) diffusion of oxygen and carbon dioxide between the alveoli and the blood, (3) transport of oxygen and carbon dioxide in the blood and body fluids to and from the cells, and (4) regulation of ventilation and other facets of respiration. The present chapter and the three following discuss, respectively, these four major aspects of respiration In subsequent chapters pulmonary disorders and special respiratory problems related to aviation medicine and deep sea diving physiology are discussed to illustrate some of the basic principles of respiratory physiology

MECHANICS OF PULMONARY VENTILATION

BASIC MECHANISMS OF LUNG EXPANSION AND CONTRACTION

The lungs can be expanded and contracted in two ways. (1) by downward and upward movement of the diaphragm to lengthen or shorten the chest cavity, and (2) by elevation and depression of the ribs to increase and decrease the anteroposterior diameter of the chest cavity. Figure 39–1 illustrates these two methods

Normal quiet breathing is accomplished almost entirely by the first of the above two methods, that is, by movement of the diaphragm During inspiration, contraction of the diaphragm pulls the lower surfaces of the lungs downward. Then, during expiration, the diaphragm simply relaxes and the *elastic recoil* of the lungs, chest wall, and abdominal structures compresses the lungs. During heavy breathing, however, the elastic forces are not powerful enough to cause the necessary rapid expiration, so that this is achieved mainly by contraction of the abdominal muscles, which forces the abdominal contents upward against the bottom of the diaphragm.

The second method for expanding the lungs to raise the rib cage. This expands the lun because, in the natural resting position, the ri slant downward, thus allowing the sternum to fa backward toward the spinal column. But, whe the rib cage is elevated, the ribs project direct forward so that the sternum now also moves fo ward away from the spine, making the anteropo terior thickness of the chest about 20 per ce greater during maximum inspiration than durir expiration. Therefore, those muscles that elevs the chest cage can be classified as muscles inspiration and those muscles that depress t chest cage, as muscles of expiration. The muscl that raise the rib cage include (1) the *sternocle domastoid* muscles that lift upward on the ste num, (2) the *anterior serrati* that lift many of t ribs, (3) the *scaleni* that lift the first two ribs, a the *external intercostals*, the function of which w be explained later. The muscles that pull the r cage downward during expiration are (1) the *a dominal recti*, which have the powerful effect pulling downward on the lower ribs at the sam

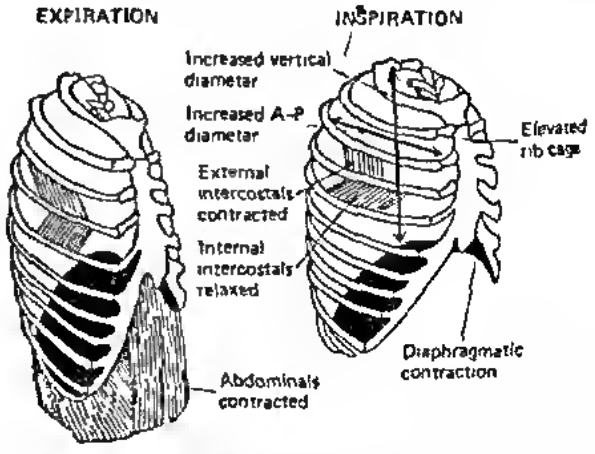


Figure 39–1. Expansion and contraction of the thoracic cage during expiration and inspiration, illustrating especially diaphragmatic contraction, elevation of the rib cage, and function of the intercostals.

time that they and the other abdominal muscles also compress the abdominal contents upward toward the diaphragm, and (2) the *internal intercostals*.

Figure 39–1 illustrates the mechanism by which the external and internal intercostals act to cause inspiration and expiration. To the left, the ribs during expiration are angled downward and the external intercostals are elongated in a forward and downward direction. As they contract, they pull the upper ribs forward in relation to the lower ribs, and this causes leverage on the ribs to raise them upward. Conversely, in the inspiratory position, the internal intercostals are elongated, and their contraction pulls the upper ribs backward in relation to the lower ribs. This causes leverage in the opposite direction and lowers the chest cage

RESPIRATORY PRESSURES

Intra-Alveolar Pressure. The respiratory muscles cause pulmonary ventilation by alternatively compressing and distending the lungs, which in turn causes the pressure in the alveoli to rise and fall. During inspiration the intra-alveolar pressure becomes slightly negative *with respect to atmospheric pressure*, normally slightly less than −1 mm Hg, and this causes air to flow inward through the respiratory passageways. During normal expiration, on the other hand, the intra-alveolar pressure rises to slightly less than +1 mm Hg, which causes air to flow outward through the respiratory passageways. Note, especially, how little pressure is required to move the air into and out of the normal lung, though many times as much pressure is required in some types of lung diseases.

During maximum expiratory effort with the glottis closed the intra-alveolar pressure can be increased to as much as 140 mm Hg in the strong, healthy man, and during maximum inspiratory effort it can be reduced to as low as −100 mm Hg.

Recoil Tendency of the Lungs, and the Intrapleural Pressure. The lungs have a continual elastic tendency to collapse and therefore to pull away from the chest wall. This is called the recoil tendency of the lungs, and it is caused by two different factors. First, throughout the lungs are many *elastic fibers* that are stretched by lung inflation and therefore attempt to shorten. Second, and even more important, the *surface tension* of the fluid lining the alveoli also causes a continual elastic tendency for the alveoli to collapse. This effect is caused by intermolecular attraction between the surface molecules of the alveolar fluid; that is, each molecule pulls on the next one so that the whole lining sheet of fluid on the alveolar surfaces acts like many small elastic balloons continually trying to collapse the lung.

Ordinarily, the elastic fibers in the lungs account for about one third of the recoil tendency, and the surface tension phenomenon accounts for about two thirds.

The total recoil tendency of the lungs can be measured by the amount of negative pressure in the pleural space required to prevent collapse of the lungs; this pressure is called the *pleural pressure* or, occasionally, the *lung recoil pressure*. It is normally about −4 mm Hg. That is, a *negative* pressure of −4 mm Hg on the outer surfaces of the lungs is required to keep them expanded to normal size. When the lungs are stretched to very large size, such as at the end of deep inspiration, the pleural pressure required then may be as great as −12 to −18 mm Hg.

"Surfactant" in the Alveoli, and Its Effect on the Collapse Tendency. A lipoprotein mixture called "surfactant" is secreted by special *surfactant-secreting cells* (the "type II granular pneumocytes") that are component parts of the alveolar epithelium. This mixture, containing especially the phospholipid *dipalmitoyl lecithin, decreases the surface tension of the fluid lining the alveoli*. In the absence of surfactant, lung expansion is extremely difficult, often requiring negative pleural pressures as low as −20 to −30 mm Hg to overcome the collapse tendency of the alveoli. This illustrates that surfactant is exceedingly important for minimizing the effect of surface tension in causing collapse of the lungs.

A few newborn babies, especially premature babies, do not secrete adequate quantities of surfactant, which makes lung expansion difficult. Without immediate and very careful treatment, most of these die soon after birth because of inadequate ventilation. This condition is called *hyaline membrane disease* or *respiratory distress syndrome*.

Surfactant acts by decreasing the forces between the surface molecules of the alveolar fluid. In the absence of surfactant, this interface has a surface tension of about 50 dynes/cm. By contrast, with *surfactant*, the *surface tension* varies *between 5* and 30 dynes/cm, averaging about four times less than in the absence of surfactant.

Role of Surfactant in "Stabilizing" the Sizes of the Alveoli. The effect of surface tension in causing collapse of an alveolus becomes very much greater as the diameter of the alveolus decreases. This effect is explained by the law of Laplace, which was discussed in Chapter 18 in relation to the blood vessels. That is, the pressure that is required to keep the alveolus expanded is directly proportional to the tension in the alveolar wall (in this instance, the surface tension of the fluid) *divided by* the diameter. Therefore, as the diameter becomes less, the pressure required to keep the alveolus from collapsing further becomes proportionately greater.

Now, let us see what would happen if two occluded alveoli were connected together and one of them were slightly smaller than the other. Obviously the pressure generated in the air in the small alveolus, because of the greater collapse tendency of the small alveolus, would be considerably greater than the pressure generated in the larger alveolus. Consequently, air would be displaced from the smaller alveolus into the larger one,

and the smaller alveolus would thus become still smaller and its pressure still greater while the larger alveolus would become larger but its pressure less. And this process would continue until the smaller alveolus would collapse entirely while displacing all of its air into the larger one. This phenomenon is called *instability of the alveoli*.

On the other hand, surfactant plays a very important role in preventing this unstable condition. Experiments with surfactant have shown that as an alveolus becomes smaller and the surfactant becomes more concentrated at the surface of the alveolar lining fluid, the surface tension becomes progressively more reduced. On the other hand, as an alveolus becomes larger and the surfactant is spread more thinly on the fluid surface, the surface tension becomes much greater. Thus, this special characteristic of surfactant helps to "stabilize" the sizes of the alveoli, causing the larger alveoli to contract more and the smaller ones to contract less. This effect obviously helps to insure that the alveoli in any one area of the lung all remain approximately the same size.

Still another factor also helps to insure that the alveoli maintain approximately equal diameters. This is the phenomenon called *interdependence*. That is, the walls of adjacent alveoli are mutually attached to each other; therefore, it is difficult physically for one alveolus to contract without the other one contracting at the same time, this also helps assure equality of alveolar size.

Role of Surfactant in Preventing Accumulation of Edema Fluid in the Alveoli. The surface tension of the fluid in the alveoli not only tends to cause collapse of the alveoli, but it also tends to pull fluid into the alveoli from the alveolar wall. In the normal lung, when there are adequate amounts of surfactant, the surface tension even then pulls fluid away from the wall with an average pressure of −3 mm Hg. Fortunately, though, there is normally an even more negative pressure in the pulmonary interstitium, averaging about −9 mm Hg, that tends to absorb fluid from the alveoli; this is what keeps the alveoli dry (see Chapter 32). On the other hand, in the absence of surfactant, the average surface tension force tending to pull fluid into the alveoli may become as great as −10 to −20 mm Hg, and this is now enough to cause massive filtration of fluid out of the alveolar wall capillaries into the alveoli, thus filling the alveoli with fluid. Therefore, one of the consequences of the absence of surfactant can be severe pulmonary edema. For instance, in the respiratory distress syndrome of the newborn, the amount of surfactant is greatly decreased, and a major number of the alveoli are filled with fluid, this fluid is one of the factors that cause the severe respiratory distress.

EXPANSIBILITY OF THE LUNGS AND THORAX: "COMPLIANCE"

The expansibility of the lungs and thorax is called *compliance*. This is expressed as the *volume increase in the lungs for each unit increase in alveolar pressure* or *for each unit decrease in pleural pressure*. The compliance of the normal lungs and thorax combined is 0.13 liter per centimeter of water pressure. That is, every time the alveolar pressure is increased by 1 cm of water, the lungs expand 130 milliliters.

Compliance of the Lungs Alone. The lungs alone, when removed from the chest, are almost twice as distensible as the lungs and thorax together, because the thoracic cage itself must also be stretched when the lungs are expanded *in situ*. Thus, the compliance of the normal lungs when removed from the thorax is about 0.22 liter per cm water. This illustrates that the muscles of inspiration must expend energy not only to expand the lungs but also to expand the thoracic cage around the lungs.

Measurement of Lung Compliance. Compliance of the lungs is measured in the following way. The person's glottis must first be completely open and remain so, so that the alveolar pressure will be equal to atmospheric pressure. Then air is inspired in steps of approximately 50 to 100 ml at a time, and pressure measurements are made at the end of each step from an intraesophageal balloon (which measures almost exactly the pleural pressure). Then the air is expired, also in steps, until the lung volume returns to the original expiratory resting level. The relationship of lung volume to pressure is then plotted as illustrated in Figure 39–2. This graph shows that the plot during inspiration is a different curva from that during expiration, which is caused by the viscous properties of the lungs. The average compliance is represented by the dashed line in the figure. Thus, in the normal-sized person the lung volume increases about 220 ml for a change in translung pressure (atmospheric pressure in the alveoli of the lung minus intraesophageal pressure) of 1 cm water. Therefore, the compliance in this instance is 0.22 liter per cm water.

Slight modifications of this procedure can be used to measure the compliance of the lungs and thoracic cage together.

Factors That Cause Abnormal Compliance. Any condition that destroys lung tissue, causes it to become fibrotic or edematous, blocks the bron-

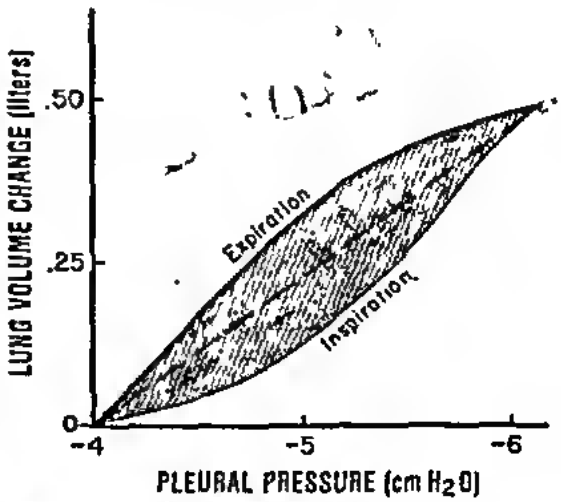


Figure 39–2. Compliance diagram in a normal person. This diagram shows the compliance of the lungs alone.

chioles, or in any other way impedes lung expansion and contraction causes decreased lung compliance. When considering the compliance of both the lung and thorax together, one must also include any abnormality that reduces the expansibility of the thoracic cage. Thus, deformities of the chest cage, such as kyphosis, severe scoliosis, and other restraining conditions such as fibrotic pleurisy or paralyzed and fibrotic muscles can all reduce the expansibility of the chest and thereby reduce the total pulmonary compliance.

THE "WORK" of BREATHING

We have already pointed out that during normal quiet respiration respiratory muscle contraction occurs only during inspiration, whereas expiration is entirely a passive process caused by elastic recoil of the lung and chest cage structures. Thus, the respiratory muscles normally perform "work" only to cause inspiration and not at all to cause expiration.

The work of inspiration can be divided into three different fractions; (1) that required to expand the lungs against its elastic forces, called *compliance work* or *elastic work*), (2) that required to overcome the viscosity of the lung and chest wall structures, called *tissue resistance work*, and (3) that required to overcome airway resistance during the movement of air into the lungs, called *airway resistance work*. These three different types of work are illustrated graphically in Figure 39–3. In this diagram the curve labeled "inspiration" illustrates the progressive *change in pleural pressure* and *lung volume* during inspiration, and the total shaded area of the figure represents the total work performed by the expiratory muscles during the act of inspiration. The shaded area, in turn, is divided into three different segments representing the three different types of work performed during inspiration. These can be explained as follows:

Compliance Work. The dotted area represents the compliance work that is required to expand the lungs against the elastic forces. This can be calculated by multiplying the volume of expansion times the average pressure required to cause the expansion, and this is equal to the area represented by the dots. That is,

$$\text{Compliance work} = \frac{\Delta V \cdot \Delta P}{2}$$

where ΔV is the increase in volume, and ΔP is the increase in intrapleural pressure.

Tissue Resistance Work. The area represented by the vertical bars is proportional to the work expenditure required to overcome the viscosity of the lungs and chest cage.

Airway Resistance Work. Finally, the area in Figure 39–3 illustrated by the horizontal bars represents the work required to overcome the resistance to airflow through the respiratory passageways.

Comparison of the Different Types of Work. It is clear from Figure 39–3 that during normal quiet breathing most of the work performed by the respiratory muscles is used simply to expand the lungs. A small amount, normally only a few per cent of the total work, is used to overcome tissue resistance (tissue viscosity), and somewhat more is used to overcome airway resistance. On the other hand, during very heavy breathing, when air must flow through the respiratory passageways at very high velocity, the greater proportion of the work is then used to overcome airway resistance.

In pulmonary disease, all three of the different types of work are frequently vastly increased. Compliance work and tissue resistance work are especially increased by diseases that cause fibrosis of the lungs, and airway resistance work is especially increased by diseases that obstruct the airways.

During normal quiet respiration, no muscle "work" is performed during expiration because expiration results from elastic recoil of the lungs and chest. But in heavy breathing or when airway resistance and tissue resistance are great, then expiratory work does of course occur and sometimes becomes even greater than inspiratory work. This is especially true in asthma, which often increases airway resistance manyfold during expiration but much less so during inspiration for reasons that will be explained later

Energy Required for Respiration. During normal quiet respiration, only 2 to 3 per cent of the total energy expended by the body is required to energize the pulmonary ventilatory process. During very heavy exercise, the absolute amount of energy required for pulmonary ventilation can increase as much as 25-fold. However, this still does not represent a significant increase in *percentage* of total energy expenditure because the total energy release in the body increases at the same time as much as 15- to 20-fold. Thus, even in heavy exercise only 3 to 5 per cent of the total energy expended is used for ventilation.

On the other hand, pulmonary diseases that decrease the pulmonary compliance, that increase airway resistance, or that increase the viscosity of the lung or chest wall can at times increase the work of breathing so much that one third or more of the total energy expended by the body is for respiration alone. Such respiratory diseases can proceed to the point that this excess work load alone is the cause of death.

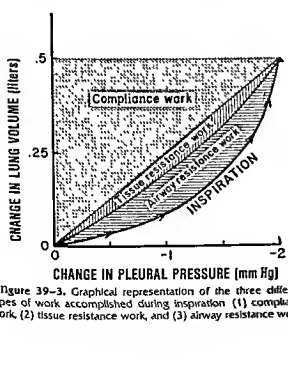


Figure 39–3. Graphical representation of the three different types of work accomplished during inspiration (1) compliance work, (2) tissue resistance work, and (3) airway resistance work.

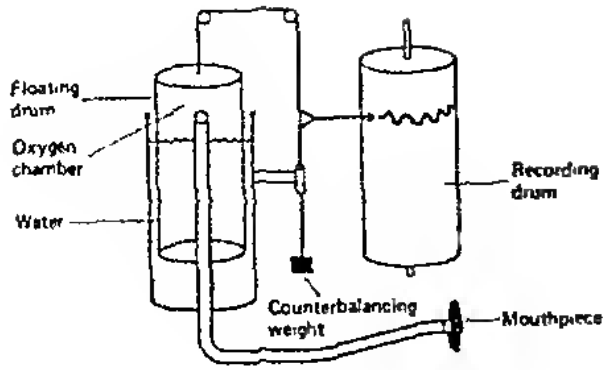


Figure 39–4. A spirometer

THE PULMONARY VOLUMES AND CAPACITIES

RECORDING CHANGES IN PULMONARY VOLUME—SPIROMETRY

A simple method for studying pulmonary ventilation is to record the volume movement of air into and out of the lungs, a process called *spirometry*. A typical spirometer is illustrated in Figure 39–4. This consists of a drum inverted over a chamber of water, with the drum *counterbalanced by a weight*. In the drum is a breathing mixture of gases, usually air or oxygen; a tube connects the mouth with the gas chamber. When one breathes in and out of the chamber the drum rises and falls, and an appropriate recording is made on a moving sheet of paper.

Figure 39–5 illustrates a spirogram showing changes in lung volume under different conditions of breathing. For ease in describing the events of pulmonary ventilation, the air in the lungs has been subdivided at different points on this diagram into four different *volumes* and four different *capacities*, which are:

THE PULMONARY "VOLUMES"

To the left in Figure 39–5 are listed four different pulmonary lung "volumes" which, when added together, equal the maximum volume to which the lungs can be expanded. The significance of each of these volumes is the following:

1. The *tidal volume* is the volume of air inspired or expired with each normal breath, and it amounts to about 500 milliliters in the average young adult man.
2. The *inspiratory reserve volume* is the extra volume of air that can be inspired over and beyond the normal tidal volume, and it is usually equal to approximately 3000 milliliters.
3. The *expiratory reserve volume* is the amount of air that can still be expired by forceful expiration after the end of a normal tidal expiration; this normally amounts to about 1100 milliliters.
4. The *residual volume* is the volume of air still remaining in the lungs after the most forceful expiration. This volume averages about 1200 milliliters.

THE PULMONARY "CAPACITIES"

In describing events in the pulmonary cycle, it *is sometimes desirable to consider two or more of* the above volumes together. Such combinations are called *pulmonary capacities*. To the right in Figure 39–5 are listed the different pulmonary capacities, which can be described as follows:

1. The *inspiratory capacity* equals the *tidal volume* plus the *inspiratory reserve volume*. This is the amount of air (about 3500 ml) that a person can breathe beginning at the normal expiratory level and distending the lungs to the maximum amount.

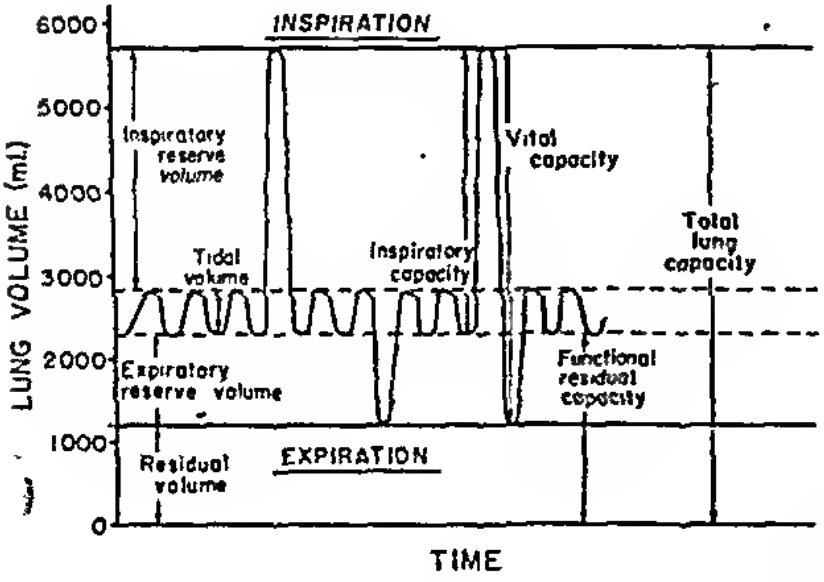


Figure 39–5. Diagram showing respiratory excursions during normal breathing and during maximal inspiration and maximal expiration.

2. The *functional residual capacity* equals the *expiratory reserve volume* plus the *residual volume*. This is the amount of air remaining in the lungs at the end of normal expiration (about 2300 ml).

3. The *vital capacity* equals the *inspiratory reserve volume* plus the *tidal volume* plus the *expiratory reserve volume*. This is the maximum amount of air that a person can expel from the lungs after first filling the lungs to their maximum extent and then expiring to the maximum extent (about 4600 ml).

4. The *total lung capacity* is the maximum volume to which the lungs can be expanded with the greatest possible inspiratory effort (about 5800 ml); it is equal to the vital capacity plus the residual volume.

All pulmonary volumes and capacities are about 20 to 25 per cent less in women than in men, and they obviously are greater in large and athletic persons than in small and asthenic persons.

Resting Expiratory Level. Normal pulmonary ventilation is accomplished almost entirely by the muscles of inspiration. On relaxation of the inspiratory muscles the elastic properties of the lungs and thorax cause the lungs to contract passively. Therefore, when all inspiratory muscles are completely relaxed the lungs return to a relaxed state called the *resting expiratory level*. The volume of air in the lungs at this level is equal to the functional residual capacity, or about 2300 milliliters in the young adult.

SIGNIFICANCE OF THE PULMONARY VOLUMES AND CAPACITIES

In normal persons the volume of air in the lungs depends primarily on body size and build. Furthermore, the different "volumes" and "capacities" change with the position of the body, most of them decreasing when the person lies down and increasing on standing. This change with position is caused by two major factors: first, a tendency for the abdominal contents to press upward against the diaphragm in the lying position, and, second, an increase in the pulmonary blood volume in the lying position, which correspondingly decreases the space available for pulmonary air.

Significance of the Residual Volume. The residual volume represents the air that cannot be removed from the lungs even by forceful expiration. This is important because it provides air in the alveoli to aerate the blood even between breaths. Were it not for the residual air, the concentrations of oxygen and carbon dioxide in the blood would rise and fall markedly with each respiration, which would certainly be disadvantageous to the respiratory process.

Significance of the Vital Capacity. Other than the anatomical build of a person, the major factors that affect vital capacity are (1) the position of the person during the vital capacity measurement, (2) the strength of the respiratory muscles, and (3) the distensibility of the lungs and chest cage, which is called "pulmonary compliance."

The average vital capacity in the young adult man is about 4.6 liters, and in the young adult woman about 3.1 liters, though these values are much greater in some persons of the same weight than in others. A tall, thin person usually has a higher vital capacity than an obese person, and a well-developed athlete may have a vital capacity as great as 30 to 40 per cent above normal—that is, 6 to 7 liters.

Vital Capacity Following Paralysis of the Respiratory Muscles. Paralysis of the respiratory muscles, which often occurs following spinal cord injuries or poliomyelitis, can cause a great decrease in vital capacity, to as low as 500 to 1000 milliliters—barely enough to maintain life. This decrease may be even lower in the case of respirator patients.

Decreased Vital Capacity Caused by Diminished Pulmonary Compliance. Obviously, any factor that reduces the ability of the lungs to expand also reduces the vital capacity. Thus, tuberculosis, chronic asthma, lung cancer, chronic bronchitis, and fibrotic pleurisy can all reduce the pulmonary compliance and thereby decrease the vital capacity. For this reason vital capacity measurements are among the most important and yet simplest of all clinical respiratory measurements for assessing the progress of different types of pulmonary fibrotic diseases.

Changes in Vital Capacity Resulting from Pulmonary Congestion. In left heart failure or any other disease that causes pulmonary vascular congestion and edema, the vital capacity becomes reduced because excess fluid in the lungs decreases lung compliance.

Vital capacity measurements made periodically in left-sided heart failure are a good means for determining whether the person's condition is deteriorating or getting better, for these measurements can indicate the degree of pulmonary edema.

ABBREVIATIONS AND SYMBOLS USED IN PULMONARY FUNCTION STUDIES

Spirometry is only one of many measurement procedures that the pulmonary physician uses daily. Some of the more important of these will be presented as we continue with our discussions of pulmonary physiology. It will become clear as we proceed that the clinical practice of respiratory medicine rests very heavily on basic physiological principles. Furthermore, we shall see that many of the measurement procedures depend heavily on mathematical computations; one such computation will be given in the following section describing the measurement of functional residual capacity. To simplify pulmonary function data, a number of abbreviations and symbols have become standardized. Some of the more important of these are given in Table 39-1. Using these symbols, we present here a few simple algebraic exercises showing some of the interrelationships among the pulmonary volumes and capacities; the student should think through and verify these interrelationships:

$$VC = IRV + VT + ERV$$
$$VC = IC + ERV$$
$$TLC = VC + RV$$
$$TLC = IC + FRC$$
$$FRC = ERV + RV$$

Table 39–1. LIST OF ABBREVIATIONS AND SYMBOLS FOR PULMONARY FUNCTION

V_T	tidal volume	P_B	atmospheric pressure
FRC	functional residual capacity	P_{alv}	alveolar pressure
ERV	expiratory reserve volume	P_{pl}	pleural pressure
RV	residual volume	P_{O_2}	partial pressure of oxygen
IC	inspiratory capacity	P_{CO_2}	partial pressure of carbon dioxide
IRV	inspiratory reserve volume	P_{N_2}	partial pressure of nitrogen
TLC	total lung capacity	$P_{a_{O_2}}$	partial pressure of oxygen in arterial blood
VC	vital capacity	$P_{a_{CO_2}}$	partial pressure of carbon dioxide in arterial blood
Raw	resistance of tracheobronchial tree to flow of air into the lung	$P_{A_{O_2}}$	partial pressure of oxygen in alveolar gas
C	compliance	$P_{A_{CO_2}}$	partial pressure of carbon dioxide in alveolar gas
V_D	volume of dead space gas	$P_{A_{H_2O}}$	partial pressure of water in alveolar gas
V_A	volume of alveolar gas	R	respiratory exchange ratio
V_I	inspired volume of ventilation per minute	Q	cardiac output
V_E	expired volume of ventilation per minute	Qs	shunt flow
V_A	alveolar ventilation per minute	$C_{a_{O_2}}$	concentration of oxygen in arterial blood
V_{O_2}	rate of oxygen uptake per minute	$C_{\bar{v}_{O_2}}$	concentration of oxygen in mixed venous blood
V_{CO_2}	amount of carbon dioxide eliminated per minute	S_{O_2}	percentage saturation of hemoglobin with oxygen
V_{CO}	rate of carbon monoxide uptake per minute	$S_{a_{O_2}}$	percentage saturation of hemoglobin with oxygen in arterial blood
$D_{L_{O_2}}$	diffusing capacity of the lung for oxygen		
$D_{L_{CO}}$	diffusing capacity of the lung for carbon monoxide		

DETERMINATION OF FUNCTIONAL RESIDUAL CAPACITY—THE HELIUM DILUTION METHOD

The functional residual capacity, which is the volume of air in the lungs at the end of normal expiration, is also the volume of air that normally remains in the lungs between breaths, which is very important to lung function. Its value changes markedly in some types of pulmonary disease, for which reason it is often desirable to measure this capacity. However, the spirometer cannot measure the residual volume of the lungs (the volume remaining after maximum expiration), and this volume composes about half the functional residual capacity. Therefore, an indirect method must be used, usually the helium dilution method.

A spirometer of known volume is filled with air mixed with helium at a known concentration. Before breathing from the spirometer, the person expires normally. At the end of this expiration the remaining volume of gases in the lungs is exactly equal to the functional residual capacity. At this point the subject immediately begins to breathe from the spirometer, and the gases of the spirometer begin to mix with the gases of the lungs. As a result, the helium becomes diluted by the functional residual capacity gases, and the volume of the functional residual capacity can then be calculated from the degree of dilution of the helium, using the following formula:

$$FRC = \left(\frac{Ci_{He}}{Cf_{He}} - 1\right) V_{I_{spir}}$$

in which

FRC is *functional residual capacity*
Ci_{He} is *initial concentration of helium in the spirometer*
Cf_{He} is *final concentration of helium in the spirometer*
$V_{I_{spir}}$ is *initial volume of the spirometer*

Once the functional residual capacity has been determined, the residual volume can then be determined by subtracting the expiratory reserve volume from the functional residual capacity. Also, the total lung capacity can be determined by adding the inspiratory capacity to the functional residual capacity. That is,

$$RV = FRC - ERV$$

and

$$TLC = FRC + IC$$

THE MINUTE RESPIRATORY VOLUME—RESPIRATORY RATE TIMES TIDAL VOLUME

The *minute respiratory volume* is the total amount of new air moved into the respiratory passages each minute, and this is equal to the tidal volume times the respiratory rate. The normal tidal volume is about 500 milliliters, and the normal respiratory rate is approximately 12 breaths per minute. Therefore, the *minute respiratory volume averages about 6 liters per minute.* A person can occasionally live for short periods of time with a minute respiratory volume as low as 1.5 liters per minute and with a respiratory rate as low as two to four breaths per minute.

The respiratory rate occasionally rises to as high as 40 to 50 per minute, and the tidal volume can become as great as the vital capacity, about 4600 milliliters in the young adult man. However, at rapid breathing rates a person usually cannot sustain a tidal volume greater than about one half the vital capacity.

ALVEOLAR VENTILATION

The ultimate importance of the pulmonary ventilatory system is to renew continually the air in

the gas exchange areas of the lungs where the air is in close proximity to the pulmonary blood. These areas include the alveoli, the alveolar sacs, the alveolar ducts, and the respiratory bronchioles. The rate at which new air does reach these areas is called *alveolar ventilation*. Strangely enough, though, during normal quiet respiration the volume of air in the tidal air is only enough to fill the respiratory passageways down as far as the terminal bronchioles, with only a very small portion of the inspired air actually flowing all the way into the alveoli. Therefore, how does new air move this last short distance from the terminal bronchioles into the alveoli? The answer is: by *diffusion*. Diffusion is caused by the kinetic motion of molecules, each gas molecule moving at high velocity among the other molecules. Fortunately, the velocity of movement of the molecules in the respiratory air is so great and the distance so short from the terminal bronchioles to the alveoli that the gases move this remaining distance in only a fraction of a second.

The Dead Space, and Its Effect on Alveolar Ventilation

Unfortunately, some of the air that a person breathes never reaches the gas exchange areas but instead goes to fill the respiratory passages. This air is called *dead-space air* because it is not useful for the gas exchange process; the respiratory passages where no gas exchange takes place is called the *dead space*.

On expiration, the air in the dead space is expired first before any of the air from the alveoli reaches the atmosphere. Therefore, the dead space is equally disadvantageous for removal of the expiratory gases from the lungs.

Measurement of the Dead Space Volume. A simple method for measuring dead space volume is illustrated by Figure 39–6. In making this measurement the subject first breathes normal air and then suddenly takes a breath of oxygen. This, obviously, fills the entire dead space with pure oxygen, and some of the oxygen also mixes with the alveolar air. Then the person expires through a rapidly recording nitrogen meter, which makes the record shown in the figure. The first portion of the expired air contains only pure oxygen, and the per cent of nitrogen is zero, but about one quarter of the way into expiration, as the alveolar air reaches the nitrogen meter, the nitrogen concentration rises, in this instance up to 60 per cent, and then levels off. The total volume of air expired in this instance was 500 milliliters, and it can readily be seen that the area covered by the dots represents the air that has no nitrogen in it. Therefore, this area also represents the dead space portion of the expired air. The area covered by the hatching represents the air containing nitrogen and therefore is the alveolar portion of the expired air. Thus, one can determine the amount of dead space air from the following equation:

$$V_D = \frac{\text{Area of dots} \times V_E}{\text{Area of hatching} + \text{Area of dots}}$$

where V_D is *dead space air*, and V_E is *total volume of expired air*.

Let us assume, for instance, that the area of the dots on the graph is 30 cm², the area of the hatching is 70 cm², and the total volume expired is 500 ml. The dead space then would be

$$\frac{30}{30 + 70} \times 500, \text{ or } 150 \text{ ml}$$

Normal Dead Space Volume. The normal dead space air in the young adult is about 150 milliliters. This increases slightly with age.

Anatomic Versus Physiologic Dead Space. The method just described for measuring the dead space measures the volume of all the space of the respiratory system besides the gas exchange areas, and this is called the *anatomic dead space*. On occasion, however, some of the alveoli themselves are not functional or are only partially functional because of absent or poor blood flow through adjacent pulmonary capillaries. Therefore, from a functional point of view, these alveoli must also be considered to be dead space. When the alveolar dead space is included in the total measurement of dead space, this is then called *physiologic dead space*, in contradistinction to the anatomic dead space. In the normal person, the anatomic and the physiologic dead spaces are nearly equal because all alveoli are functional in the normal lung, but in persons with partially functional or nonfunctional alveoli in some parts of the lungs, the physiologic dead space is sometimes as much as ten times the anatomic dead space, or as much as 1 to 2 liters. These problems will be discussed further in Chapter 40 in relation to pulmonary gaseous exchange and in Chapter 43 in relation to certain pulmonary diseases.

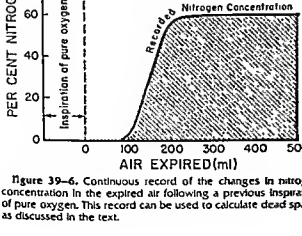


Figure 39–6. Continuous record of the changes in nitrogen concentration in the expired air following a previous inspiration of pure oxygen. This record can be used to calculate dead space, as discussed in the text.

RATE OF ALVEOLAR VENTILATION

Alveolar ventilation per minute is the total volume of new air entering the alveoli (other adjacent gas exchange areas).

equal to the respiratory rate times the amount of new air that enters the alveoli with each breath:

$$\dot{V}_A = \text{Freq} \cdot (V_T - V_D)$$

where V_A is the *volume of alveolar ventilation per minute*, Freq is the *frequency of respiration per minute*, V_T is the *tidal volume*, and V_D is the *dead space volume*.

Thus, with a normal tidal volume of 500 milliliters, a normal dead space of 150 milliliters, and a respiratory rate of 12 times per minute, alveolar ventilation equals 12 × (500 − 150), or 4200 milliliters per minute.

Theoretically, when the tidal volume falls to equal the dead space volume, no new air at all should enter the alveoli with each breath, and the alveolar ventilation per minute should become zero however rapidly the person breathes. However, this is not entirely true because the flow patterns in the passageways allow some alveolar air to be expired before all the dead space air is expired. Also, the same is true for inspiration. Therefore, there can be a slight amount of alveolar ventilation even with tidal volumes of as little as 60 to 75 milliliters.

On the other hand, when the tidal volume is several liters, the effect of dead space volume on alveolar ventilation obviously becomes almost insignificant.

Alveolar ventilation is one of the major factors determining the concentrations of oxygen and carbon dioxide in the alveoli. Therefore, almost all discussions of gaseous exchange in the following chapters emphasize alveolar ventilation. *The respiratory rate, the tidal volume, and the minute respiratory volume are mainly of importance only insofar as they affect alveolar ventilation.*

FUNCTIONS OF THE RESPIRATORY PASSAGEWAYS

THE TRACHEA, BRONCHI, AND BRONCHIOLES

Figure 39–7 illustrates the respiratory system, showing especially the respiratory passageways. After passing through the nose and pharynx, the air is distributed to the lungs by way of the trachea, bronchi, and bronchioles. The *trachea* is called the *first generation respiratory passageway*, and the two *main, right* and *left bronchi* are the *second generation*; then each division thereafter is an additional generation. There are between 20 and 25 generations before the air finally reaches the alveoli. The final few generations of passageways, after the passageways become smaller than 1 to 1.5 mm in diameter, are called *bronchioles*; all the intermediate generations between the trachea and the bronchioles are the *bronchi*.

One of the most important problems in all the respiratory passageways is to keep them open to allow easy passage of the air to and from the alveoli. To keep the trachea from collapsing, multiple cartilage rings extend about five sixths of the way around the trachea. In the walls of the bronchi are less extensive cartilage plates that also maintain a reasonable amount of rigidity yet allow sufficient motion for the lungs to expand and contract. These plates become progressively less extensive in the later generations of bronchi and are completely gone in the bronchioles. The bronchioles, on the other hand, are not prevented from collapsing by any rigidity of their walls. Instead, they are expanded by the same transpulmonary pressures that expand the

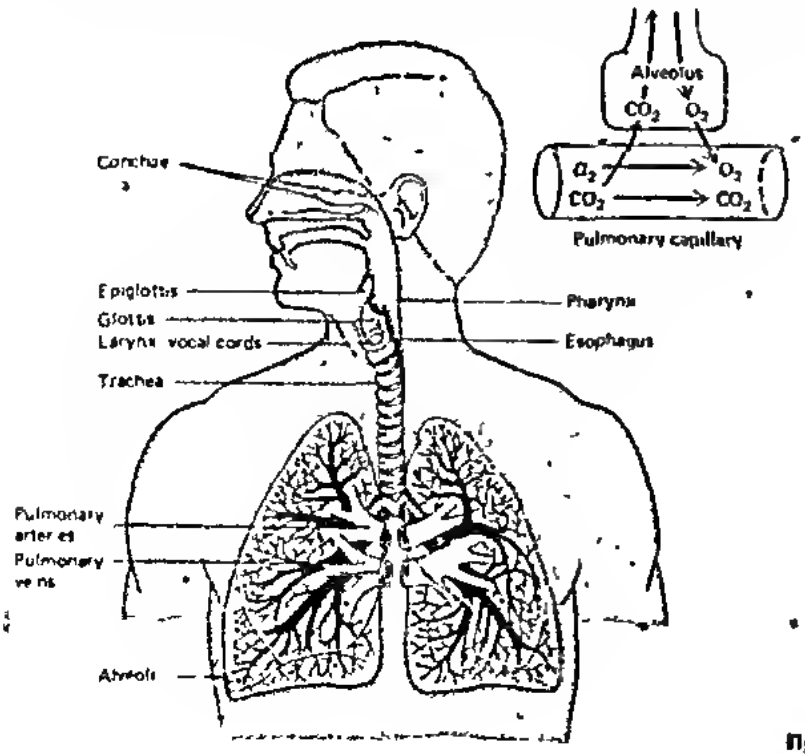


Figure 39–7. The respiratory passages

alveoli. That is, as the alveoli enlarge, so also do the bronchioles enlarge.

The Muscular Wall of the Bronchioles and Its Control. In all areas of the trachea and bronchi not occupied by cartilage plates, the walls are composed mainly of smooth muscle. The walls of the bronchioles are almost entirely smooth muscle, with the exception of the most terminal bronchiole, called the *respiratory bronchiole*, that has only a few smooth muscle fibers. Many obstructive diseases of the lung cause narrowing of the muscular bronchioles, often because of excessive contraction of the smooth muscle itself. Therefore, it has become quite important to understand those factors that control bronchiolar contraction.

Under normal respiratory conditions, the bronchioles remain mainly in a relaxed condition, and air flows easily through these so that less than a 1 mm Hg air pressure gradient from the alveoli to the trachea is sufficient to cause all the air flow required for quiet breathing. But this is not true during disease. For instance, in asthma the degree of contraction is sometimes so great that 20 times as much pressure may be required to provide adequate air flow; this is especially true during expiration, as we shall see later.

Nervous Control of the Bronchioles. The only important nervous control of the bronchioles is by way of parasympathetic nerve fibers carried in the vagus nerves. These nerves secrete acetylcholine and when activated cause mild-to-moderate constriction of the bronchioles. Though it is doubtful that parasympathetic stimulation by itself ever causes serious bronchiolar constriction, nevertheless, when disease processes have already caused some constriction, superimposed nervous stimulation often seems to worsen the condition. In these conditions, administration of drugs that block the effects of acetylcholine, such as the drug *atropine*, often can relax the bronchioles enough to relieve the obstruction.

Humoral Control of the Bronchi. Several different humoral substances are often quite active in causing bronchiolar constriction. Two of the most important of these are *histamine* and the substance called *slow reactive substance of anaphylaxis*. Both of these are released in the lung tissues by mast cells during allergic reactions, especially allergic reactions to pollen breathed in the air. Therefore, they play key roles in causing the airway obstruction that occurs in allergic asthma; this is especially true of the slow reactive substance of anaphylaxis. Also, irritants entering the airways—such as smoke, dust, sulfur dioxide, and some of the acidic elements in smog—can all initiate similar local reactions that cause obstructive constriction of the bronchioles.

In contrast to the humoral substances that constrict the bronchioles, two other hormones, epinephrine and norepinephrine, both of which are secreted by the adrenal glands in response to sympathetic stimulation, relax the bronchioles. Therefore, activation of the sympathetic nervous system is often valuable in relaxing the airways and preventing obstruction. For instance, this occurs when the sympathetic nervous system is activated during heavy exercise. In addition, injection of one of these two substances intravenously is often of value in treating severe respiratory obstruction.

The Mucous Coat of the Respiratory Passageways, and Action of Cilia to Clear the Passageways. All the respiratory passages, from the nose to the terminal bronchioles, are kept moist by a layer of mucus that coats the entire surface. This mucus is secreted partly by individual goblet cells in the epithelial lining of the passages and partly by small submucous glands. In addition to keeping the surfaces moist, the mucus also traps small particles out of the inspired air and keeps most of these from ever reaching the alveoli. Then, the mucus itself is removed from the passages in the following manner:

The entire surface of the respiratory passages, both in the nose and in the lower passages, is lined with ciliated epithelium, with about 200 cilia on each epithelial cell. These beat continually at a rate of 10 to 20 times per second by the mechanism explained in Chapter 2, and the direction of their "power stroke" is always toward the pharynx. That is, the cilia in the lower respiratory passages beat upward while those in the nose beat downward. This continual beating causes the coat of mucus to flow slowly, at a velocity of about 1 cm per minute, toward the pharynx. Then the mucus and its entrapped particles are either swallowed or coughed to the exterior.

The Cough Reflex

The bronchi and the trachea are so sensitive to light touch that excessive amounts of any foreign matter or any other cause of irritation initiates the cough reflex. The larynx and carina (the point where the trachea divides into the bronchi) are especially sensitive, and the terminal bronchioles and alveoli are very sensitive to corrosive chemical stimuli such as sulfur dioxide gas and chlorine. Afferent impulses pass from the respiratory passages mainly through the vagus nerves to the medulla. There, an automatic sequence of events is triggered by the neuronal circuits of the medulla, causing the following effects:

First, about 2.5 liters of air is inspired. Second, the epiglottis closes, and the vocal cords shut tightly to entrap the air within the lungs. Third, the abdominal muscles contract forcefully, pushing against the diaphragm while other expiratory muscles, such as the internal intercostals, also contract forcefully. Consequently, the pressure in the lungs rises to as high as 100 or more mm Hg. Fourth, the vocal cords and the epiglottis suddenly open widely so that air under pressure in the lungs *explodes* outward. Indeed, this air is sometimes expelled at velocities as high as 75 to 100 miles an hour. Furthermore, and very important, the strong compression of the lungs also collapses the bronchi and trachea by causing the noncartilaginous parts of these to invaginate inward so that the exploding air actually passes through *bronchial* and *tracheal clefts*. The rapidly moving air usually carries with it any foreign matter that is present in the bronchi or trachea.

The Sneeze Reflex

The sneeze reflex is very much like the cough reflex except that it applies to the nasal passageways instead of to the lower respiratory passages. The initiating stimulus of the sneeze reflex is irritation in the nasal passageways, the afferent impulses passing in the fifth nerve to the medulla where the reflex is triggered. A series of reactions similar to those for the cough reflex takes place; however, the uvula is depressed so that large amounts of air pass rapidly through the nose, as well as through the mouth, thus helping clear the nasal passages of foreign matter.

MAXIMUM EXPIRATORY FLOW

In many respiratory diseases, especially in asthma, the resistance to air flow becomes especially great during expiration, sometimes causing tremendous difficulty in breathing For this reason, a decided attempt has been made to understand the factors that have unique effects in limiting a person's ability to expire air from the lungs This has led to the concept called the *maximum expiratory flow*, which can be defined as follows When a person expires with great force, the expiratory air flow reaches a maximum flow beyond which the flow cannot be increased even with greatly increased additional force This is the maximum expiratory flow. However, the maximum expiratory flow is much greater when the lungs are filled with a large volume or air than when they are almost empty. These principles can be understood by referring to Figure 39–8.

Section A of Figure 39–8 illustrates the effect of pressure applied to the outsidea of the alveoli and the passageways caused by compressing the chest cage The arrows on both the outsides of the alveoli and the outsides of the respiratory bronchioles indicate that the same amount of pressura is applied to tha outsides of both the alveoli and the bronchioles Therefore, not only does this pressure force air from the alveoli into the bronchioles, but it tends also to rollapse the bronchioles at the same time, thus greatly increasing the airway resistanca and opposing the movement of the air to the exterior. That is, the expiratory force is opposed by increased resistance. Once the bronchioles have berome almost rompletely rollapsed, further expiratory force can still greatly increase the alveolar pressure, but it also increases the airway resistance by an equal amount, thus preventing any further increase in flow Therefore, beyond a critical degree of expiratory force, the maximum expiratory flow has been reached.

Section B of Figure 39–8 illustrates the effect of bronchiolar rollapse of the maximum expiratory flow. The curve recorded in this section shows the maximum

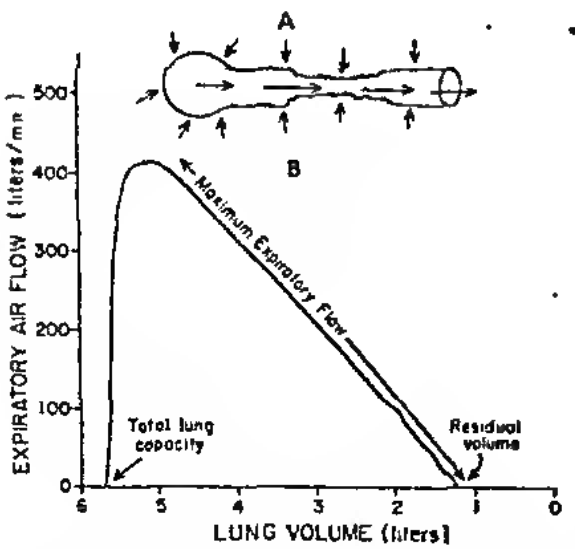


Figure 39–8. *A*, Collapse of the respiratory passageway during maximum expiratory effort, an effect that limits the expiratory flow rate. *B*, Effect of lung volume on the maximum expiratory air flow, showing decreasing maximum expiratory air flow as the lung volume becomes smaller

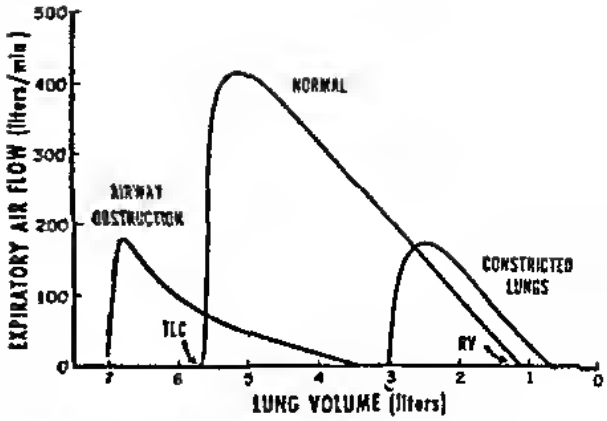


Figure 39–9. Effect of two different respiratory abnormalities—constricted lungs and airway obstruction—on the maximum expiratory flow-volume curve

expiratory flow achieved by a normal person who fir inhales as much air as possible and then expires wi maximum expiratory effort until he can expire no mor Note that ha quickly reaches a *maximum expiratory a flow* of over 400 liters/min. But it does not matter ho much additional expiretory effort he exerts, thia ia sti the maximum flow that he can achieve.

Note also that as the lung volume becomes smalli thia maximum expiratory flow also becomes less Tl main reason for this is that in the enlarged lung tl bronchi and bronchioles are held open partially v elastic pull on their outsides by lung atructurol el ments; however, as the lung becomes smaller, the structures are relaxed so that tha bronchi and bro chioles collapse more essily.

Abnormalities of the Maximum Expiratory Flov Volume Curve. The maximum expiretory flow-volum curve is often recorded in the pulmonary function lab ratory to determina abnormalities of pulmonary vent lation Figure 39–9 illustrates onre mora the norm curve and curves recorded in two different types of lun diseases (1) constricted lungs and (2) airway obstru tion

Note that the *constricted lungs* have both reduce totaf lung capacity (TLC) and reduced residual volum (RV). Furthermore, sinre the lung cannot expand to it normal volume, even with the greatest possible expira tory effort the maximal expiretory flow cannot rise t equal that of the normal curve Constricted lung disease include fibrotic diseasea of the lung itself such as *tuber culosis, silicosis,* and others, and also diseases tha constrict the chest cage such as *kyphosis, scoliosis,* an *fibrotic pleurisy.*

In diseases with *airway obstruction*, it is usually muc more difficult to expire than to inspire because th closing tendency of the airways is greatly increased b the positive pressure in the chest during expiratior while the negative pleural pressure of inspiration ac tually "pulls" the airways open at the same time that i expands the alveoli. Therefore, air tends to enter th lung easily but then become trapped in the lungs Ove a period of months or years, this effect increases botl the total lung capacity and the residual volume, a illustrated in Figure 39–9. Also, because of the obstruc tion of the airways and also because these collapse mor easily than normal airways, the maximum expirator flow is greatly reduced. The classic disease that cause

severe airway obstruction is *asthma*. However, serious airway obstruction also occurs in some stages of *emphysema*.

FORCED EXPIRATORY VITAL CAPACITY AND FORCED EXPIRATORY VOLUME

Another exceedingly useful clinical pulmonary test, and one that is also very simple, is to make a record on a spirometer of the *forced expiratory vital capacity* (FVC). Such a record is illustrated in Figure 39–10A for a person with normal lungs and in Figure 39–10B for a person with airway obstruction. In performing the forced expiratory vital capacity maneuver the person first inspires maximally to the total lung capacity, then exhales into the spirometer with maximum expiratory effort as rapidly and as completely as possible. The total excursion of the record represents the "forced vital capacity," as illustrated in the figure.

Now, study the difference between the two records, for normal lungs and airway obstruction. The forced vital capacities are nearly equal, indicating only moderate difference in basic lung volumes in the two persons. On the other hand, there is a major difference in the maximum flow rate at which the persons can expire, especially during the first second. Therefore, it is customary to record the forced expiratory volume during the first second (FEV_1) and to use this for comparison between the normal and the abnormal. In the normal person the percentage of the forced vital capacity that is expired in the first second ($FEV_1/FVC\%$) is about 80. However, note in Figure 39–10B that with airway obstruction this value was decreased to only 47 per cent. In serious airway obstruction, as often occurs in acute asthma, this can decrease to less than 20 per cent.

RESPIRATORY FUNCTIONS OF THE NOSE

As air passes through the nose, three distinct functions are performed by the nasal cavities. First, the air

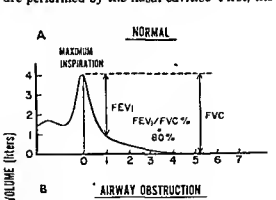


Figure 39–10. Recordings during a forced vital capacity maneuver. *A*, in the normal person, and *B*, in the person with airway obstruction.

is warmed by the extensive surfaces of the conchae and septum, a total area of about 160 cm², which are illustrated in Figure 39–7. Second, the *air is almost completely humidified* even before it passes beyond the nose. Third, the *air is filtered*. All these functions together are called the *air conditioning function* of the upper respiratory passageways. Ordinarily, the air rises to within 2 to 3 per cent of body temperature and within 2 to 3 per cent of full saturation with water vapor before it reaches the trachea. When a person breathes air through a tube directly into the trachea (as through a tracheostomy), the cooling and, especially, the drying effect in the lower lung can lead to serious lung crusting and infection.

Filtration Function of the Nose. The hairs at the entrance to the nostrils are important for the removal of large particles. Much more important, though, is the removal of particles by *turbulent precipitation*. That is, the air passing through the nasal passageways hits many obstructing vanes, the conchae (also called "turbinates" because they cause turbulence of the air), the septum, and the pharyngeal wall. Each time air hits one of these obstructions it must change its direction of movement, the particles suspended in the air, having far more mass and momentum than air, cannot change their direction of travel as rapidly as can the air. Therefore, they continue forward, striking the surfaces of the obstructions where they become entrapped in the mucous coating and are transported by the cilia to the pharynx and swallowed.

Size of Particles Entrapped in the Respiratory Passages. The nasal turbulence mechanism for removing particles from air is so effective that almost no particles larger than 4 to 6 microns in diameter enter the lungs through the nose. This size is smaller than the size of red blood cells. Consequently, by far the greater proportion of dust and other large particles in air is removed before the air finally reaches the lungs.

Of the remaining particles, many of sizes between 1 and 5 microns *settle* out in the smaller bronchioles as a result of *gravitational precipitation*. For instance, terminal bronchiolar disease is very common in coal miners because of settled dust particles. Some of the still smaller particles (smaller than 1 micron in diameter) *diffuse* against the walls of the alveoli and adhere to the alveolar fluid. But many particles smaller than 0.5 micron in diameter remain suspended in the alveolar air and are later expelled by expiration. For instance, the particles of cigarette smoke have a particle size of approximately 0.3 micron. Almost none of these are precipitated in the respiratory passageways before they reach the alveoli. However, up to one third of them do precipitate in the alveoli by the diffusion process, with the balance remaining suspended and expelled in the expired air.

Particles that become entrapped in the alveoli are removed mainly by *alveolar macrophages*, as explained in Chapter 5. An excess of particles causes growth of fibrous tissue in the alveolar septa, leading to permanent debility.

VOCALIZATION

Speech involves the respiratory system particularly, but it also involves (1) specific speech nervous control centers in the cerebral cortex, which will be discussed in Chapter 55, (2) respiratory control centers of the

brain, and (3) the articulation and resonance structures of the mouth and nasal cavities. Basically, speech is composed of two separate mechanical functions: (1) *phonation*, which is achieved by the larynx, and (2) *articulation*, which is achieved by the structures of the mouth

Phonation. The larynx is specially adapted to act as a vibrator. The vibrating element is the *vocal cords*, which are folds along the lateral walls of the larynx that are stretched and positioned by several specific muscles within the confines of the larynx itself

Figure 39–11A illustrates the basic structure of the larynx, showing that each vocal cord is stretched between the *thyroid cartilage* anteriorly and an *arytenoid cartilage* posteriorly The specific muscles within the larynx that position and control the degree of stretch of the vocal cords are also shown Thus, one can see in the figure that contraction of the *posterior cricoarytenoid* muscles pulls the arytenoid cartilages away from the thyroid cartilage and thereby stretches the vocal cords. The *transverse arytenoid muscle* pulls the arytenoid cartilages together and, therefore, approximates the two vocal cords so that they vibrate in a stream of expired air Conversely, contraction of the *lateral cricoarytenoid muscles* pulls tha arytenoid cartilages forward and apart to allow normsl respiration

The *thyroarytenoid muscles* lie within the folds of the vocal cords themselves They are made up of many small slips of muscle controlled separately by different nerve fibers Tho slips of muscle adjacent to the edgea of the vocal cord can contract separately from those adjacent to the wsll of the larynx, and other individual portions of these muscles can also contract independently of each other These contractions control the *shope* of the vocal cords—whether thick or thin, or with sharp or blunt edges—and they also help control tha tension of the cords during different types of phonstion

Vibration of the Vocal Cords. One might suspect that the vocal cords would vibrate in the direction of the flowing air However, this is not the case Instead, they vibrate laterally The cause of tha vibration is the following When the vocal cords are approximated and air is expired, pressure of the air from below first pushes the vocal cords apart, which allows rapid flow of air between their margins The rapid flow of air then immediately creates a partial vacuum between the vocal cords because of the Bernoulli principle, which pulls them once again toward each other. This stops the flow of air, pressure builds up behind the cords, and the cords open once more, thus continuing in a vibratory pattern.

Frequency of Vibration. The pitch of the sound emitted by the larynx can be changed in two different ways. First, a change can be achieved by *stretching* or *relaxing the vocal cords*. The mechanisms involved were partly explained above in the discussion of the intrinsic laryngeal muscles; but, in addition to the effects of the intrinsic muscles, the muscles attached to the external surfaces of the larynx can also pull against the cartilages and thereby help to stretch or relax the vocal cords. For instance, the entire larynx is moved upward by the external laryngeal muscles, which helps to stretch the vocal corda when one wishes to emit a very high frequency sound, and the larynx is moved downward, with corresponding loosening of the vocal cords, when one wishes to emit a very bass sound

The second means for changing the sound frequency is *to change the shape and mass of the vocal card edges* When very high frequency sounds are emitted, different slips of the thyroarytenoid muscles contract in such a way that the edges of the vocal cords are sharpened and thinned, whereas when bass frequencies are emitted, the thyroarytenoid muscles contract in a different pattern so that broad edges with a large mass are approximated. Figure 39–11B shows some of the positions and shapes of the vocal cords during different types of phonation.

Articulation and Resonance. The three major organs of articulation are the *lips*, the *tongue*, and the *soft palate* These need not be discussed in detail because all of us are familiar with their movements during speech and other vocalizations.

The resonators include the *mouth*, the *nose* and *associated nasal sinuses*, the *pharynx*, and even the *chest cavity* itself Here again we are all familiar with the resonating qualities of these different structures. For instance, the function of the nasal resonators is illustrated by the change in quality of the voice when a person has a severe cold that blocks the air passages to these resonators

ARTIFICIAL RESPIRATION

MOUTH-TO-MOUTH BREATHING

A very successful method of artificial respiration is mouth-to-mouth breathing, in which the operator rapidly inspires a deep breath and then breathes into the mouth of the subject This method has often been shunned in the past because of the belief that the expired

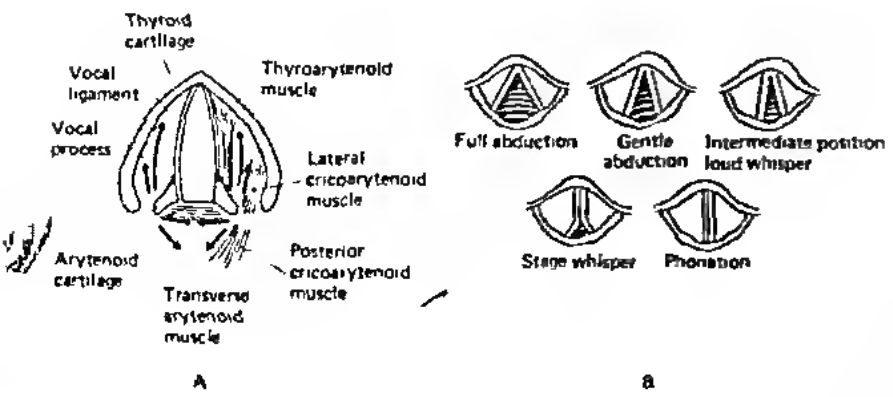


Figure 39–11. Laryngeal function in phonation. (Modified from Greene The Voice and Its Disorders. 4th ed Philadelphia, J B Lippincott Company, 1980)

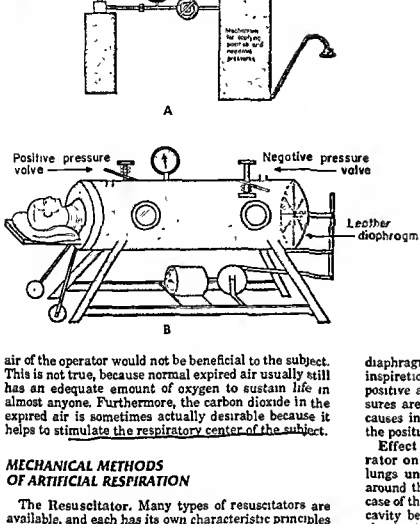


Figure 39–12. A, The resuscitator. B, Tank respirator.

air of the operator would not be beneficial to the subject. This is not true, because normal expired air usually still has an adequate amount of oxygen to sustain life in almost anyone. Furthermore, the carbon dioxide in the expired air is sometimes actually desirable because it helps to stimulate the respiratory center of the subject.

MECHANICAL METHODS OF ARTIFICIAL RESPIRATION

The Resuscitator. Many types of resuscitators are available, and each has its own characteristic principles of operation. Basically, the resuscitator, illustrated in Figure 39–12A, consists of a supply of oxygen or air, a mechanism for applying intermittent positive pressure and, with some machines, negative pressure as well, and a mask that fits over the face of the patient. This apparatus forces air through the mask into the lungs of the patient during the positive pressure cycle and then either allows the air to flow passively out of the lungs during the remainder of the cycle or pulls the air out by negative pressure.

Earlier resuscitators often caused such severe damage to the lungs because of excessive positive pressure that their usage was at one time greatly decried. However, most resuscitators now have safety valves that prevent the positive pressure from rising usually above +14 mm Hg and the negative pressure from falling below −9 mm Hg. These pressure limits are adequate to cause excellent artificial respiration of *normal lungs* and yet are slight enough to prevent damage.

The Tank Respirator. Figure 39–12B illustrates the usual tank respirator with a patient's body inside the tank and his head protruding through a flexible but airtight collar. At the end of the tank opposite the patient's head is a motor-driven leather diaphragm that moves back and forth with sufficient excursion to raise and lower the pressure inside the tank. As the leather diaphragm moves inward, positive pressure develops around the body and causes expiration; and as the diaphragm moves outward, negative pressure causes inspiration. Check valves on the respirator control the positive and negative pressure. Ordinarily these pressures are adjusted so that the negative pressure that causes inspiration falls to −10 to −20 cm water end the positive pressure rises to 0 to +5 cm water.

Effect of the Resuscitator and the Tank Respirator on Venous Return. When air is forced into the lungs under positive pressure, or when the pressure around the patient's body is greatly reduced, as in the case of the tank respirator, the pressure inside the chest cavity becomes greater than the pressure everywhere else in the body. Therefore, the flow of blood into the chest from the peripheral veins becomes impeded. As a result, use of excessive positive pressures with either the resuscitator or the tank respirator can reduce the cardiac output—sometimes to lethal levels. A person can usually survive as much as 20 mm Hg continuous positive pressure in the lungs, but exposure for more than a few minutes to greater than 30 mm Hg may cause death.

REFERENCES

Agostoni, E.: Thickness and pressure of the pleural liquid. *In* Fishman, A. P., and Hecht, H. H. (eds.): The Pulmonary Circulation and Interstitial Space. Chicago, University of Chicago Press, 1969, p. 65.

Bless, D. M. (ed.): Vocal Fold Physiology: Contemporary Research and Clinical Issues. San Diego, College-Hill Press, 1983.

Bradley, G. W.: Control of the breathing pattern. *Int. Rev. Physiol.*, *14* 185, 1977.

Bryant, C.: The Biology of Respiration. Baltimore, University Park Press, 1979.

Burri, P. H.: Fetal and postnatal development of the lung. *Annu. Rev. Physiol.*, *46* 617, 1984.

Comroe, J. H., Jr., et al.: The Lung: Clinical Physiology and Pulmonary Function Tests. 2nd Ed. Chicago, Year Book Medical Publishers, 1963.

Culver, B. H., and Butler, J.: Mechanical influences on the pulmonary microcirculation. *Annu. Rev. Physiol.*, *42* 187, 1980.

Cumming, G.: Alveolar ventilation. *Recent model analysis. In* MTP International Review of Science. Physiology Vol 2. Baltimore, University Park Press, 1974, p. 139.

Drazen, J. M., et al.: High-frequency ventilation. *Physiol. Rev.*, *64* 505, 1984.

Ellis P D, and Billings, D M Cardiopulmonary Resuscitation Procedures for Basic and Advanced Life Support. St. Louis C V Mosby, 1979

Engel, L. A, and Macklem, P T Gas mixing and distribution in the lung *Int Rev Physiol*, *14* 37, 1977

Feldman, S A and Crawley, B E (eds) Tracheostomy and Artificial Ventilation in the Treatment of Respiratory Failure Baltimore Williams & Wilkins, 1977

Fink, B R The Human Larynx A Functional Study New York, Raven Press, 1975

Fishman, A P Assessment of Pulmonary Function New York, McGraw Hill, 1980

Forster, R. E Pulmonary ventilation and blood gas exchange *In* Sodeman, W A Jr, and Sodeman W A (eds) Pathologic Physiology Mechanisms of Disease, 5th Ed Philadelphia W B Saunders Co, 1974 p 371

Gil J Organization of microcirculation in the lung *Annu Rev Physiol*, *42* 177, 1980

Green M and Moxham, J The respiratory muscles. *Clin Sci*, *68* 1, 1985

Greene M The Voice and Its Disorders, 2nd Ed Philadelphia J B Lippincott Co, 1965

Guyton A C Electronic counting and size determination of particles in aerosols *J Indust Hyg Toxicol* *28* 133, 1946

Guyton, A C Measurement of the respiratory volumes of laboratory animals *Am J Physiol* *150* 70 1947

Guyton, A C Analysis of respiratory patterns in laboratory animals *Am J Physiol* *150* 78 1947

Herada, R N and Repine J E Pulmonary host defense mechanisms *Chest* *87* 247 1985

Harding R. Function of the larynx in the fetus and newborn *Annu Rev Physiol* *46* 645, 1984

Healy G B, and McGill T J I (eds) Laryngo-Tracheal Problems in the Pediatric Patient Springfield, Ill Charles C Thomas 1979

Hyatt, R E and Black L F The flow volume curve A current perspective *Am Rev Resp Dis* *107* 191, 1973

Jarvis J F An Introduction to the Anatomy and Physiology of Speech and Hearing Cape Town South Africa Juta & Co 1978

Kao F F An Introduction to Respiratory Physiology New York, American Elsevier Publishing Co, 1972

Levitzky M G Pulmonary Physiology New York McGraw-Hill Book Co, 1982

Luchsinger R and Arnold G Voice-Speech Language Clinical Communicology—Its Physiology and Pathology Belmont, Cal, Wadsworth Publishing Co 1965

Macklem P T Airway obstruction and collateral ventilation *Physiol Rev* *51* 368 1971

Macklem, P T.: Respiratory mechanics *Annu Rev Physiol*, *40* 15 1978

Mead J Respiration Pulmonary mechanics *Annu Rev Physio* *35* 169, 1973

Murray, J F The Normal Lung 2nd ed Philadelphia, W. B. Saunde Co, 1986

Nadel J A, *et al* Control of mucus secretion and ion transport airways *Annu Rev Physiol*, *41* 369, 1979

Nagaishi, C Functional Anatomy and Histology of the Lung Baltimo University Park Press, 1973

Notter, R. H, and Finkelstein, J N Pulmonary surfactant An inte disciplinary approach *J Appl Physiol*, *57* 1613, 1984

Phillipson, E A. Respiratory adaptations in sleep. *Annu. Rev Physio* *40* 133, 1978

Rahn, H, *et al* The pressure-volume diagram of the thorax and lun *Am J Physiol*, *146* 161, 1946

Ryan J W Processing of endogenous polypeptides by the lungs. Ann *Rev Physiol*, *44* 241, 1982

Scarpelli, E, and Auld, P A M (eds) Pediatric Pulmonary Physiolog Philadelphia, Lea & Febiger, 1975

Scheich, O C H, and Schreiner, C (eds). Hearing Mechanisms a Speech New York, Springer-Verlag, 1979

Schneiderman, C R. Basic Anatomy and Physiology in Speech a Hearing San Diego, College-Hill Press, 1984

Shearer W M Illustrated Speech Anatomy Springfield, Ill., Charl C Thomas, 1978

Singh, R P Anatomy of Hearing and Speech New York, Oxfo University Press, 1980

Strang, L. B. Fetal and newborn lung *In* MTP International Review of Science Physiology Vol 2 Baltimore, University Park Press, 1974 p 31

Stuart, B O Deposition of inhaled aerosols *Arch Intern Med*, *131* 60, 1973

Thurlbeck, W M Structure of the lungs. *Int Rev Physiol*, *14* 1, 1977

Tierney, D F Lung metabolism and biochemistry *Annu Rev Physiol*, *36* 209, 1974

Wagner, P D Ventilation perfusion relationships. *Annu Rev Physiol*, *42* 235, 1980

Weibel, E R and Backofen, H Structural design of the alveolar septum and fluid exchange *In* Fishman, A P, and Renkin, E M (eds) Pulmonary Edema Baltimore, Waverly Press, 1979 p 1

West, J B Respiratory Physiology Baltimore, Williams & Wilkins, 1974

White, F N Comparative aspects of vertebrate cardiorespiratory physiology *Annu Rev Physiol*, *40* 471, 1978

Wyman R J Neural generation of the breathing rhythm *Annu Rev Physiol*, *39* 417, 1977

40

Physical Principles of Gaseous Exchange; Diffusion of Oxygen and Carbon Dioxide Through the Respiratory Membrane

After the alveoli are ventilated with fresh air, the next step in the respiratory process is *diffusion* of oxygen from the alveoli into the pulmonary blood and diffusion of carbon dioxide in the opposite direction—from the pulmonary blood into the alveoli. The process of diffusion is simple, involving merely random molecular motion of molecules, these intertwining their ways in both directions through the respiratory membrane. However, in respiratory physiology one is concerned not only with the basic mechanism by which diffusion occurs but also with the *rate* at which it occurs, and *this is a much more complicated problem, requiring* a rather deep understanding of the physics of diffusion and gaseous exchange. Therefore, a brief review of this subject is presented here as a prelude to the main text of this chapter.

PHYSICS OF DIFFUSION AND GAS PRESSURES

THE MOLECULAR BASIS OF GASEOUS DIFFUSION

All the gases that are of concern in respiratory physiology are simple molecules that are free to move among each other, which is the process called "diffusion." This is also true of the gases dissolved in the fluids and tissues of the body.

However, for diffusion to occur, there must be a source of energy. This is provided by the kinetic motion of the molecules themselves. That is, except at absolute zero temperature, all molecules of all matter are continually undergoing some type of motion. For free molecules that are not physically attached to others, this means linear movement at high velocity of the molecules until they strike another molecule. Then they bounce away in new directions and continue again until striking still other molecules. In this way the molecules move rapidly among each other.

Net Diffusion of a Gas in One Direction—Effect of a Concentration Gradient. However, if a gas chamber or a solution has a high concentration of a gas at one end of the chamber and a low concentration at the other end, as illustrated in Figure 40–1, net diffusion of the gas will occur from the high concentration area toward the low area. The reason for this is obvious: There simply are far more molecules at end A of the chamber to diffuse toward end B than there are molecules to diffuse in the opposite direction. Therefore, the rates of diffusion in each of the two directions are proportionately different, as illustrated by the lengths of the two arrows, the *net* diffusion rate, which is the difference between the rates in the two directions, is represented by the difference between the lengths of the two arrows.

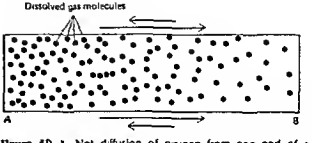


Figure 40–1. Net diffusion of oxygen from one end of a chamber to the other

GAS PRESSURES IN A MIXTURE OF GASES—PARTIAL PRESSURES OF INDIVIDUAL GASES

Pressure is caused by the constant impaction of kinetically moving molecules against a surface. Therefore, the pressure of a gas acting on the surfaces of the respiratory passages and alveoli is proportional to the summated force of impaction of all the molecules striking the surface at any given instant. And this means that *the total pressure is directly proportional to the concentration of the gas molecules.*

However, in respiratory physiology, one deals with mixtures of gases, mainly of oxygen, nitrogen, and carbon dioxide. Furthermore, the rate of diffusion of each of these gases is directly proportional to the pressure caused by this gas alone, which is called the *partial pressure* of the respective gas. Therefore, let us explain the concept of partial pressure.

Consider air which has an approximate composition of 79 per cent nitrogen and 21 per cent oxygen. The total pressure of this mixture is 760 mm Hg, and it is clear from the above description of the molecular basis of pressure that each gas contributes to the total pressure in direct proportion to its relative concentration. Therefore, 79 per cent of the 760 mm Hg is caused by nitrogen (about 600 mm Hg) and 21 per cent by oxygen (about 160 mm Hg). Thus, the "partial pressure" of nitrogen in the mixture is 600 mm Hg, and the "partial pressure" of oxygen is 160 mm Hg; the total pressure is 760 mm Hg, the sum of the individual partial pressures.

The partial pressures of the individual gases in a mixture are designated by the terms P_{O_2}, P_{CO_2}, P_{N_2}, P_{H_2O}, P_{He}, and so forth.

PRESSURES OF GASES IN WATER AND TISSUES

Gases dissolved in water or in the body tissues also exert pressure, because the dissolved molecules are moving randomly and have kinetic energy as do the molecules in the gaseous phase. Furthermore, when the molecules of a dissolved gas encounter a surface like the membrane of a cell, the pressure of the gas against the surface is directly proportional to the concentration of the dissolved molecules, in the same way that a gas in the gaseous phase exerts its own individual partial pressure. The pressures of the separate dissolved gases are designated similarly as for the partial pressures of the gases in the gaseous state, i.e., P_{O_2}, P_{CO_2}, P_{N_2}, P_{He}.

Factors That Determine the Concentration of a Gas Dissolved in a Fluid. The concentration of a gas in a solution is determined not only by its pressure but also by the *solubility coefficient* of the gas. That is, some types of molecules, especially carbon dioxide, are physically or chemically attracted to water molecules while others are repelled. Obviously, when molecules are attracted, far more of them can then become dissolved without building up excess pressure within the solution. On the other hand, those that are repelled will develop excessive pressures for very little degree of solubility.

Thus, since both pressure and solubility coefficient determine the volume of gas that will be dissolved in a given volume of fluid—which is the concentration of the dissolved gas—this can be expressed by the following formula, which is *Henry's law*:

$$\text{Concentration of dissolved gas} = \text{pressure} \times \text{solubility coefficient}$$

When concentration is expressed in volume of gas dissolved in each volume of water and pressure is expressed in atmospheres, the solubility coefficients for important respiratory gases at body temperature are the following.

Oxygen	0.024
Carbon dioxide	0.57
Carbon monoxide	0.018
Nitrogen	0.012
Helium	0.008

From this table, one can see that carbon dioxide is more than 20 times as soluble as oxygen, and oxygen is more soluble than the other three gases. These different solubilities are important because the solubility times the pressure of a gas determines the quantity of the gas that becomes physically dissolved in the fluids of the body, this in turn is one of the factors that determines the rate at which the gas can diffuse through the tissues.

Diffusion of Gases Between the Gaseous Phase in the Alveoli and the Dissolved Phase in the Pulmonary Blood. In the alveoli, it is necessary for the respiratory gases to move between the gaseous phase in the alveoli and the dissolved phase in the pulmonary blood. This occurs in the following way:

The partial pressure of each gas in the respiratory gas mixture tends to force molecules of that gas into solution. On the other hand, in the pulmonary fluids, the dissolved molecules of the same gas are randomly bouncing against the interface between the fluid and the gaseous phase, and some of these molecules escape back into the alveoli. The rate at which they escape is directly proportional to their partial pressure in the solution. But, in which direction will net diffusion of the gas occur? The answer to this is, net diffusion is determined by the difference between the two partial pressures. If the partial pressure is greater in the gaseous phase, then more molecules will go into solution than in the other direction. This is the direction in which oxygen moves, thus entering the pulmonary blood and then being carried throughout the body. Alternately, if the pressure of the gas is greater in the dissolved phase, then net diffusion will occur toward the gaseous phase in the alveoli, which is what happens for the movement of carbon dioxide in the lungs.

THE VAPOR PRESSURE OF WATER

When air enters the respiratory passageways, water immediately evaporates from the surfaces of these passages and humidifies the air. This results from the fact that water molecules, like the different dissolved gas molecules, are continually escaping from the water surface into the gas phase. The pressure that the water molecules exert to escape through the surface is called the *vapor pressure* of the water. At normal body temperature, 37°C, this vapor pressure is 47 mm Hg. Therefore, once the gas mixture has become fully humidified—that is, once it is in "equilibrium" with the surrounding water—the partial pressure of the water vapor in the gas mixture is also 47 mm Hg. This partial pressure, like the other partial pressures, is designated P_{H_2O}.

The vapor pressure of water depends entirely on the temperature of the water. The greater the temperature, the greater the kinetic activity of the molecules, and therefore the greater the likelihood that the water molecules will escape from the surface of the water into the gas phase. Consequently, water vapor pressure increases as the temperature increases. For instance, the water vapor pressure at 0°C is 5 mm Hg and at 100°C is 760 mm Hg. But the most important value to remember is the vapor pressure at body temperature—47 mm Hg; this value will appear in many of our subsequent discussions.

DIFFUSION OF GASES THROUGH FLUIDS—THE PRESSURE GRADIENT FOR DIFFUSION

Now, let us return to the problem of diffusion. From the above discussion it is already clear that when the concentration, or pressure, of a gas is greater in one area than in another area, there will be net diffusion from the high pressure area toward the low pressure area. For instance, returning to Figure 40–1 one can readily see that the molecules in the area of high pressure, because of their greater number, have more statistical chance of moving randomly into the area of low pressure than do molecules attempting to go in the other direction. However, some molecules do bounce from the area of low pressure toward the area of high pressure. Therefore, the net diffusion of gas from the area of high pressure to the area of low pressure is equal to the number of molecules bouncing in this direction *minus* the number bouncing in the opposite direction, and this in turn is proportional to the gas pressure difference between the two areas. The pressure in area A of Figure 40–1 minus the pressure in area B divided by the distance of diffusion is known as the *pressure gradient for diffusion* or simply the *diffusion gradient*. The rate of net gas diffusion from area A to area B is directly proportional to this gradient.

The principle of diffusion from an area of high pressure to an area of low pressure holds true for diffusion of gases in a gaseous mixture, diffusion of dissolved gases in a solution, and even diffusion of gases from the gaseous phase into the dissolved state in liquids. That is, *there is always net diffusion from areas of high pressure to areas of low pressure.*

As more and more gas diffuses from area A to area B in Figure 40–1, the pressure in area A falls while that in area B rises so that the two pressures approach each other; as a result, the net rate of diffusion becomes less and less. After a reasonable length of time, the gaseous pressures in both ends of the chamber become *essentially* equal, and, thereafter, no net diffusion of gas occurs from one end to the other end. This does not mean that no molecules of gas diffuse, but merely that as many molecules then diffuse in one direction as the other.

Quantifying the Net Rate of Diffusion in Fluids. In addition to the pressure difference, several other factors affect the rate of gas diffusion in a fluid. These are (1) the solubility of the gas in the fluid, (2) the cross-sectional area of the fluid, (3) the distance through which the gas must diffuse, (4) the molecular weight of the gas, and (5) the temperature of the fluid. In the body, the last of these factors, the temperature, remains reasonably constant and usually need not be considered.

Obviously, the greater the solubility of the gas, the greater will be the number of molecules available to diffuse for any given pressure difference. Also, the greater the cross-sectional area of the chamber, the greater will be the total number of molecules to diffuse. On the other hand, the greater the distance that the molecules must diffuse, the longer will it take the molecules to diffuse the entire distance. Finally, the greater the velocity of kinetic movement of the molecules, which at any given temperature is inversely proportional to the square root of the molecular weight, the greater is the rate of diffusion of the gas. All of these factors can be expressed in a single formula, as follows:

$$D \propto \frac{\Delta P \times A \times S}{d \times \sqrt{MW}}$$

in which D is the diffusion rate, ΔP is the pressure difference between the two ends of the diffusion pathway, A is the cross-sectional area of the pathway, S is the solubility of the gas, d is the distance of diffusion, and MW is the molecular weight of the gas.

It is obvious from this formula that the characteristics of the gas itself determine two factors of the formula: solubility and molecular weight, and these together are called the *diffusion coefficient* of the gas. That is, the diffusion coefficient equals $S/\sqrt{MW}$; the relative rates at which different gases at the same pressure levels will diffuse are proportional to their diffusion coefficients. Considering the diffusion coefficient for oxygen to be 1, the *relative* diffusion coefficients for different gases of respiratory importance in the body fluids are:

Oxygen	1.0
Carbon dioxide	20.3
Carbon monoxide	0.81
Nitrogen	0.53
Helium	0.95

DIFFUSION OF GASES THROUGH TISSUES

The gases that are of respiratory importance are highly soluble in lipids and, consequently, are also highly soluble in cell membranes. Because of this, these gases diffuse through the cell membranes with very little impediment. Instead, the major limitation to the movement of gases in tissues is the rate at which the gases can diffuse through the tissue water instead of through the cell membranes. Therefore, diffusion of gases through the tissues, including through the respiratory membrane, is almost equal to the diffusion of gases through water, as given in the above list.

COMPOSITION OF ALVEOLAR AIR—ITS RELATION TO ATMOSPHERIC AIR

Alveolar air does not have the same concentrations of gases as atmospheric air by any means, which can readily be seen by comparing the alveolar air composition in column 5 of Table 40–1 with the composition of atmospheric air in column 1. There are several reasons for the differences. First, the alveolar air is only partially replaced by atmospheric air with each breath. Second, oxygen is

Table 40–1 PARTIAL PRESSURES OF RESPIRATORY GASES AS THEY ENTER AND LEAVE THE LUNGS (AT SEA LEVEL)

	Atmospheric Air* (mm Hg)		Humidified Air (mm Hg)		Alveolar Air (mm Hg)		Expired Air (mm Hg)	
N_2	597.0	(78.62%)	563.4	(74.09%)	569.0	(74.9%)	566.0	(74.5%)
O_2	159.0	(20.84%)	149.3	(19.67%)	104.0	(13.6%)	120.0	(15.7%)
CO_2	0.3	(0.04%)	0.3	(0.04%)	40.0	(5.3%)	27.0	(3.6%)
H_2O	3.7	(0.50%)	47.0	(6.20%)	47.0	(6.2%)	47.0	(6.2%)
TOTAL	760.0	(100.0%)	760.0	(100.0%)	760.0	(100.0%)	760.0	(100.0%)

*On an average cool clear day

constantly being absorbed from the alveolar air. Third, carbon dioxide is constantly diffusing from the pulmonary blood into the alveoli. And, fourth, dry atmospheric air that enters the respiratory passages is humidified even before it reaches the alveoli.

Humidification of the Air As It Enters the Respiratory Passages. Column 1 of Table 40–1 shows that atmospheric air is composed almost entirely of nitrogen and oxygen; it normally contains almost no carbon dioxide and little water vapor. However, as soon as the atmospheric air enters the respiratory passages, it is exposed to the fluids covering the respiratory surfaces. Even before the air enters the alveoli, it becomes totally humidified.

The partial pressure of water vapor at normal body temperature of 37°C is 47 mm Hg, which, therefore, is the partial pressure of water in the alveolar air. Since the total pressure in the alveoli cannot rise to more than the atmospheric pressure, this water vapor simply *dilutes* all the other gases in the inspired air. In column 3 of Table 40–1 it can be seen that humidification of the air dilutes the oxygen partial pressure at sea level from an average of 159 mm Hg in atmospheric air to 149 mm Hg in the humidified air, and it has diluted the nitrogen partial pressure from 597 to 563 mm Hg.

RATE AT WHICH ALVEOLAR AIR IS RENEWED BY ATMOSPHERIC AIR

In the preceding chapter it was pointed out that the *functional residual capacity* of the lungs, which is the amount of air remaining in the lungs at the end of normal expiration, measures approximately 2300 ml. Furthermore, only 350 ml of new air is brought into the alveoli with each normal respiration, and the same amount of old alveolar air is expired. Therefore, the amount of alveolar air replaced by new atmospheric air with each breath is only one seventh of the total, so that many breaths are required to exchange most of the alveolar air. Figure 40–2 illustrates this slow rate of renewal of the alveolar air. In the first alveolus of the figure an excess amount of a gas has been placed momentarily in all the alveoli, but note that even at the end of 16 breaths the excess gas still has not been completely removed from the alveoli.

Figure 40–3 illustrates graphically the rate at which an excess of a gas in the alveoli is normally removed, showing that with normal alveolar ventilation approximately half the gas is removed in 17 seconds. When a person's rate of alveolar ventilation is only half normal, half the gas is removed in 34 seconds, and, when the rate of ventilation is two times normal, half is removed in about 8 seconds.

Importance of the Slow Replacement of Alveolar Air. This slow replacement of alveolar air is of particular importance in preventing sudden changes in gaseous concentrations in the blood. This makes the respiratory control mechanism much more stable than it would otherwise be and helps to prevent excessive increases and decreases in tissue oxygenation, tissue carbon dioxide concentration, and tissue pH when respiration is temporarily interrupted.

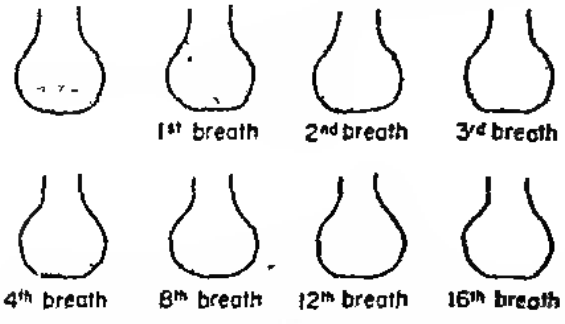


Figure 40–2. Expiration of a gaseous excess from the alveoli with successive breaths

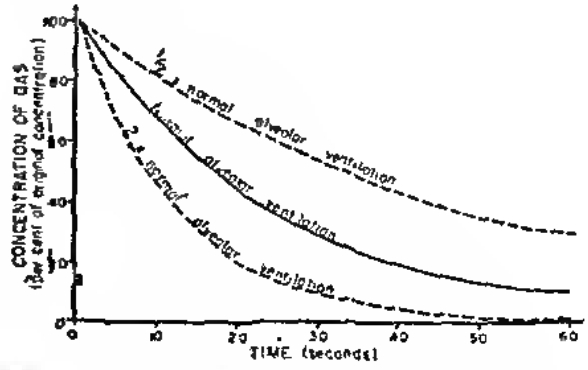


Figure 40–3. Rate of removal of excess gas from the alveoli

OXYGEN CONCENTRATION AND PARTIAL PRESSURE IN THE ALVEOLI

Oxygen is continually being absorbed into the blood of the lungs, and new oxygen is continually entering the alveoli from the atmosphere. The more rapidly oxygen is absorbed, the lower becomes its concentration in the alveoli; on the other hand, the more rapidly new oxygen is brought into the alveoli from the atmosphere, the higher becomes its concentration. Therefore, oxygen concentration in the alveoli, and therefore its partial pressure as well, is controlled by, first, the rate of absorption of oxygen into the blood and, second, the rate of entry of new oxygen into the lungs by the ventilatory process.

Figure 40–4 illustrates the effect both of alveolar ventilation and of rate of oxygen absorption into the blood on the alveolar partial pressure of oxygen ($P_{A_{O_2}}$). The solid curve represents oxygen absorption at a rate of 250 ml per minute, and the dotted curve at 1000 ml per minute. At a normal ventilatory rate of 4 2 liters per minute and an oxygen consumption of 250 ml per minute, the normal operating point in Figure 40–4 is point A. The figure also shows that when 1000 ml of oxygen is being absorbed each minute, as occurs during moderate exercise, the rate of alveolar ventilation must increasa fourfold to maintain the alveolar P_{O_2} at the normal value of 104 mm Hg, and at still higher rates of oxygen absorption the rate of ventilation must rise proportionately to maintain normal alveolar P_{O_2}.

Another effect illustrated in Figure 40–4 is that an extremely marked increase in alveolar ventilation can never increase the alveolar P_{O_2} above 149 mm Hg as long as the person is breathing normal atmospheric air, for this is the maximum content of oxygen in humidified air. However, if the person breathes gases containing concentrations of oxygen higher than 149 mm Hg, the alveolar P_{O_2} can approach these higher concentrations at very high rates of ventilation

CO_2 CONCENTRATION AND PARTIAL PRESSURE IN THE ALVEOLI

Carbon dioxide is continually being formed in the body, then discharged into the alveoli; and it is continually being removed from the alveoli by the process of ventilation. Therefore, the two factors that determine alveolar concentration of carbon dioxide, and also its partial pressure ($P_{A_{CO_2}}$), are (1) the rate of excretion of carbon dioxide from the blood into the alveoli, and (2) the rate at which carbon dioxide is removed from the alveoli by alveolar ventilation.

Figure 40–5 illustrates the effects on the alveolar P_{CO_2} of both alveolar ventilation and the rate of carbon dioxide excretion. The solid curve represents a normal rate of carbon dioxide excretion of 200 ml per minute. At the normal rate of

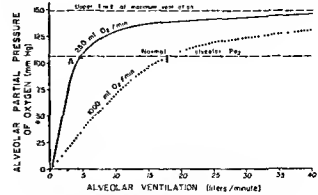


Figure 40–4. Effect of alveolar ventilation and of rate of oxygen absorption from the alveoli on the alveolar P_{O_2}

alveolar ventilation of 4.2 liters per minute, the operating point for alveolar P_{CO_2} is at point A in Figure 40–5—that is, 40 mm Hg.

Two other facts are also evident from Figure 40–5: First, *the alveolar* P_{CO_2} *increases directly in proportion to the rate of carbon dioxide excretion,* as represented by the elevation of the dotted curve for 800 ml CO_2 excretion per minute. Second, *the alveolar* P_{CO_2} *decreases in inverse proportion to alveolar ventilation.* Therefore, the concentrations and partial pressures of both oxygen and carbon dioxide in the alveoli are determined by the rates of absorption or excretion of the two gases, and also by the alveolar ventilation.

EXPIRED AIR

Expired air is a combination of dead space air and alveolar air, and its overall composition is, therefore, determined by, first, the proportion of the expired air that is dead space air and the proportion that is alveolar air Figure 40–6 shows the progressive changes in oxygen and carbon dioxide partial pressures in the expired air during the course of expiration The very first portion of this air, the dead space air, is typical humidified air as shown in column 3 of Table 40–1. Then, progressively more and more alveolar air becomes mixed with the dead space air until all the dead space air has finally

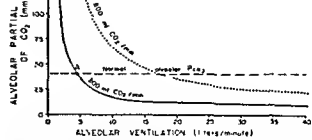


Figure 40–5. Effect on alveolar P_{CO_2} of alveolar ventilation and rate of carbon dioxide excretion from the blood.

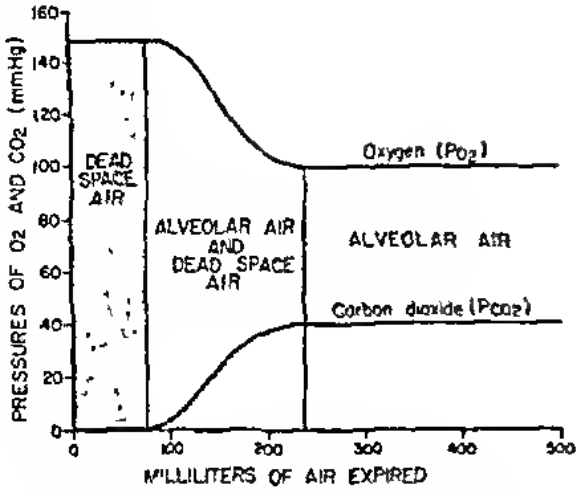


Figure 40–6. Oxygen and carbon dioxide partial pressures in the various portions of normal expired air

been washed out so that nothing but alveolar air is expired at the end of expiration. Therefore, the means for collecting alveolar air for study is simply to collect a sample of the last portion of expired air.

Normal expired air, containing both dead space air and alveolar air, has gaseous concentrations and partial pressures approximately as shown in column 7 of Table 40–1—that is, concentrations somewhere between those of alveolar air and humidified atmospheric air.

DIFFUSION OF GASES THROUGH THE RESPIRATORY MEMBRANE

The Respiratory Unit. Figure 40–7 illustrates the respiratory unit, which is composed of a *respiratory bronchiole, alveolar ducts, atria,* and *alveoli* (of which there are about 300 million in the two lungs, each alveolus having an average diameter of about 0.2 mm). The alveolar walls are extremely thin, and within them is an almost solid network of interconnecting capillaries, as illustrated in Figure 40–8. Indeed, because of the extensiveness of the capillary plexus, the flow of blood in the alveolar wall has been described as a "sheet" of flowing blood. Thus, it is obvious that the alveolar gases are in close proximity to the blood of the capillaries. Consequently, gaseous exchange between the alveolar air and the pulmonary blood occurs through the membranes of all the terminal portions of the lungs. These membranes are collectively known as the *respiratory membrane,* also called the *pulmonary membrane.*

The Respiratory Membrane. Figure 40–9 illustrates the ultrastructure of the respiratory membrane shown in cross-section. It also shows the diffusion of oxygen from the alveolus into the red blood cell and diffusion of the carbon dioxide in the opposite direction. Note the following different layers of the respiratory membrane:

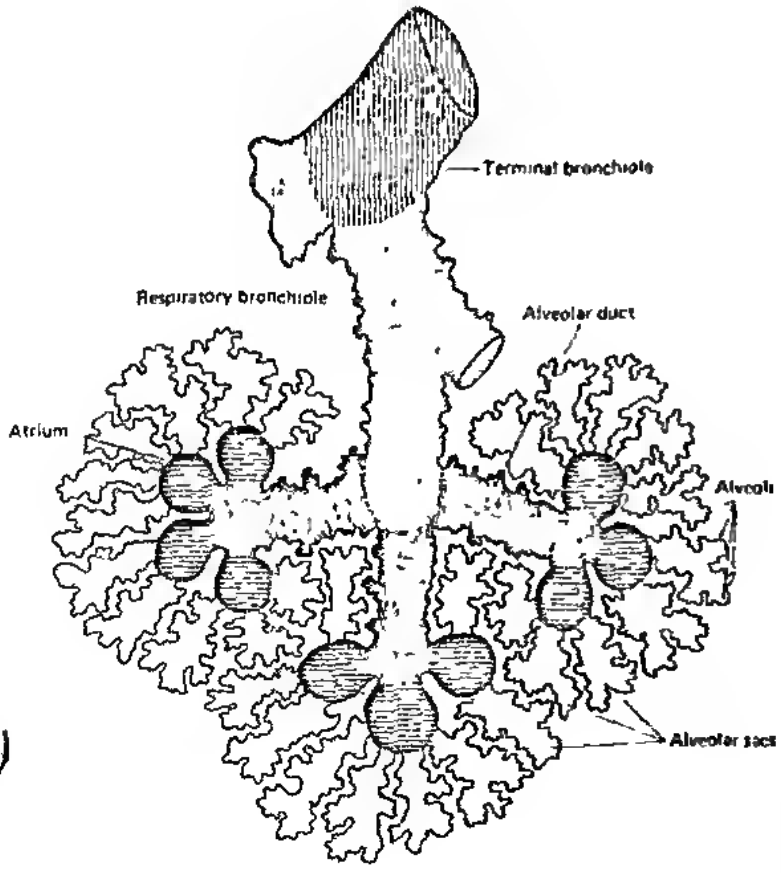
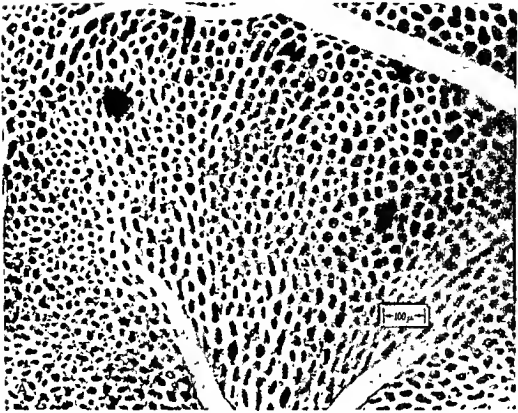
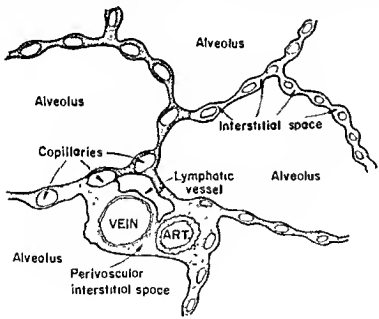


Figure 40–7. The respiratory lobule (from Miller: The Lung. Charles C Thomas.)





B

Figure 40–8. *A*, Surface view of capillaries in the alveolar wall (from Maloney and Castle *Resp. Physiol.* 7:150, 1969. Reproduced by permission of ASP Biological and Medical Press North-Holland Division.) *B*, Cross-sectional view of alveolar walls and their vascular supply

1. A layer of fluid lining the alveolus and containing surfactant that reduces the surface tension of the alveolar fluid.
2. The alveolar epithelium comprised of very thin epithelial cells.
3. An epithelial basement membrane.
4. A very thin interstitial space between the alveolar epithelium and capillary membrane.
5. A capillary basement membrane that in many places fuses with the epithelial basement membrane.
6. The capillary endothelial membrane.

Despite the large number of layers, the overall thickness of the respiratory membrane in some areas is as little as 0.2 micron, and it averages 0.63 micron.

From histologic studies it has been estimated that the total surface area of the respiratory membrane is approximately 160 square meters in the normal adult. This is equivalent to the floor area of a room 50 feet long by 30 feet wide. The total quantity of blood in the capillaries of the lung at

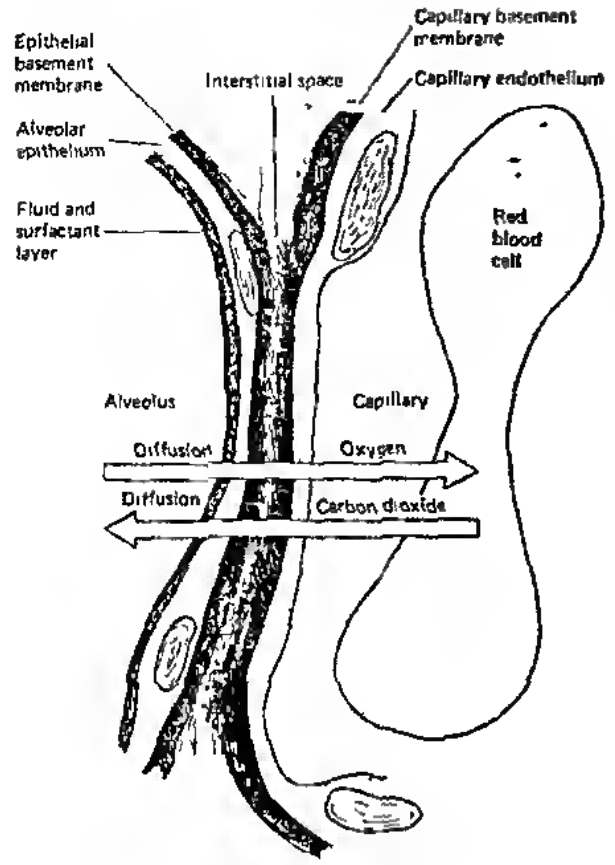


Figure 40–9. Ultrastructure of the respiratory membrane as shown in cross-section

any given instant is 60 to 140 ml. Now imagine this small amount of blood spread over the entire surface of a 50 by 30 foot floor, and it is very easy to understand the rapidity of respiratory exchange of gases.

The average diameter of the pulmonary capillaries is less than 8 microns, which means that red blood cells must actually squeeze through them. Therefore, the red blood cell membrane usually touches the capillary wall so that oxygen and carbon dioxide need not pass through significant amounts of plasma as they diffuse between the alveolus and the red cell. Obviously, this, too, increases the rapidity of diffusion.

FACTORS THAT AFFECT RATE OF GAS DIFFUSION THROUGH THE RESPIRATORY MEMBRANE

Referring to the earlier discussion of diffusion through water, one can apply the same principles and same formula to diffusion of gases through the respiratory membrane. Thus, the factors that [illegible] how rapidly a gas will pass through the [illegible] are (1) the *thickness of the membrane*, (2) the *surface area of the membrane*, (3) the *diffusion coefficient* of the gas in the substance of the membrane—that is, in the water of the membrane, and (4) the *pressure difference* between the sides of the membrane.

The *thickness of the respiratory membrane* occasionally increases, often as a result of edema fluid in the interstitial space of the membrane and in the alveoli, so that the respiratory gases must diffuse not only through the membrane but also through this fluid. Also, some pulmonary diseases cause fibrosis of the lungs, which can increase the thickness of some portions of the respiratory membrane. Because the rate of diffusion through the membrane is inversely proportional to the thickness of the membrane, any factor that increases the thickness to more than two to three times normal can interfere significantly with normal respiratory exchange of gases.

The *surface area of the respiratory membrane* may be greatly decreased by many different conditions. For instance, removal of an entire lung decreases the surface area to half normal. Also, in *emphysema* many of the alveoli coalesce, with dissolution of many alveolar walls. Therefore, the new chambers are much larger than the original alveoli, but the total surface area of the respiratory membrane is considerably decreased because of loss of the alveolar walls. When the total surface area is decreased to approximately one third to one fourth normal, exchange of gases through the membrane is impeded to a significant degree even *under resting conditions*. And, during competitive sports and other strenuous exercise, even the slightest decrease in surface area of the lungs can be a serious detriment to respiratory exchange of gases.

The *diffusion coefficient* for the transfer of each gas through the respiratory membrane depends on its *solubility* in the membrane and inversely on the *square root of its molecular weight*. The rate of diffusion in the respiratory membrane is almost exactly the same as that in water, for reasons explained above. Therefore, for a given pressure difference, carbon dioxide diffuses through the membrane about 20 times as rapidly as oxygen. Oxygen in turn diffuses about two times as rapidly as nitrogen.

The *pressure difference* across the respiratory membrane is the difference between the partial pressure of the gas in the alveoli and the pressure of the gas in the blood. The partial pressure represents a measure of the total number of molecules of a particular gas striking a unit area of the alveolar surface of the membrane in unit time, and the pressure of the gas in the blood represents the number of molecules attempting to escape from the blood in the opposite direction. Therefore, the difference between these two pressures is a measure of the *net tendency* for the gas to move through the membrane. Obviously, when the partial pressure of a gas in the alveoli is greater than the pressure of the gas in the blood, as is true for oxygen, net diffusion from the alveoli into the

blood occurs, but, when the pressure of the gas in the blood is greater than the partial pressure in the alveoli, as is true for carbon dioxide, net diffusion from the blood into the alveoli occurs.

DIFFUSING CAPACITY OF THE RESPIRATORY MEMBRANE

The ability of the respiratory membrane to exchange a gas between the alveoli and the pulmonary blood can be expressed in quantitative terms by its *diffusing capacity*, which is defined as the *volume of a gas that diffuses through the membrane each minute for a pressure difference of 1 mm Hg.*

Obviously, all the factors discussed above that affect diffusion through the respiratory membrane can affect the diffusing capacity.

The Diffusing Capacity for Oxygen. In the average young male adult the *diffusing capacity for oxygen* under resting conditions averages *21 ml per minute per* mm *Hg*. In functional terms, what does this mean? The mean oxygen pressure difference across the respiratory membrane during normal, quiet breathing is approximately 11 mm Hg. Multiplication of this pressure by the diffusing capacity (11 × 21) gives a total of about 230 ml of oxygen diffusing through the respiratory membrane each minute, and this is equal to the rate at which the body uses oxygen.

Change in Oxygen-Diffusing Capacity During Exercise. During strenuous exercise, or during other conditions that greatly increase pulmonary blood flow and alveolar ventilation, the diffusing capacity for oxygen increases in young male adults to a maximum of about 65 ml per minute per mm Hg, which is three times the diffusing capacity under resting conditions. This increase is caused by several different factors, among which are (1) opening up of a number of previously dormant pulmonary capillaries, thereby increasing the surface area of the blood into which the oxygen can diffuse, and (2) dilatation of all the pulmonary capillaries that were already open, thereby further increasing the surface area. Therefore, during exercise, the oxygenation of the blood is increased not only by increased alveolar ventilation but also by a greater capacity of the respiratory membrane for transmitting oxygen into the blood.

Diffusing Capacity for Carbon Dioxide. The diffusing capacity for carbon dioxide has never been measured because of the following technical difficulty: Carbon dioxide diffuses through the respiratory membrane so rapidly that the average P_{CO_2} in the pulmonary blood is not far different from the P_{CO_2} in the alveoli—the average difference is less than 1 mm Hg—and with the available techniques, this difference is too small to be measured.

Nevertheless, measurements of diffusion of other gases have shown that the diffusing capacity varies directly with the diffusion coefficient of the particular gas. Since the diffusion coefficient of carbon dioxide is 20 times that of oxygen, one would expect a diffusing capacity for carbon dioxide under resting conditions of about 400 to 450 ml and during exercise of about 1200 to 1300 ml per minute per mm Hg.

The importance of these high diffusing capacities for carbon dioxide is this: When the respiratory membrane becomes progressively damaged, its capacity for transmitting oxygen into the blood is often impaired enough to cause death of the person while carbon dioxide diffusion can still occur in adequate amounts. However, a condition in which low diffusing capacity for carbon dioxide does cause significant difficulty is when lung damage is far beyond that which ordinarily causes death but the person's life is being maintained by intensive oxygen therapy that overcomes the reduction in oxygen-diffusing capacity; unfortunately oxygen therapy does not help the carbon dioxide diffusion.

Figure 40–10 compares the measured or calculated diffusing capacities of oxygen, carbon dioxide, and carbon monoxide at rest and during exercise, showing the extreme diffusing capacity of carbon dioxide and also the effect of exercise on the diffusing capacities of all the gases.

Measurement of Diffusing Capacity. The oxygen-diffusing capacity can be calculated from measurements of (1) alveolar P_{O_2}, (2) P_{O_2} in the pulmonary capillary blood, and (3) the rate of oxygen uptake by the blood. However, to measure the P_{O_2} in the pulmonary capillary blood is so difficult and so imprecise that it is not practical to measure oxygen-diffusing capacity by such a direct procedure, except on an experimental basis.

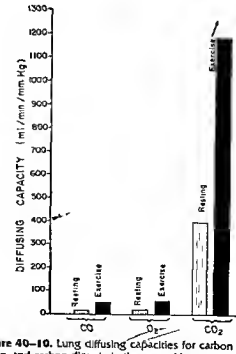


Figure 40–10. Lung diffusing capacities for carbon monoxide, oxygen, and carbon dioxide in the normal lungs.

To obviate the difficulties encountered in measuring oxygen-diffusing capacity, physiologists usually measure carbon monoxide-diffusing capacity instead, and then calculate the oxygen-diffusing capacity from this. The principle of the carbon monoxide method is the following: A small amount of carbon monoxide is breathed into the alveoli, and the partial pressure of the carbon monoxide in the alveoli is measured from appropriate alveolar air samples. Under most conditions it can be assumed that the carbon monoxide pressure in the blood is essentially zero because hemoglobin combines with this gas so rapidly that its pressure never has time to build up. Therefore, the pressure difference of carbon monoxide across the respiratory membrane is equal to its partial pressure in the alveoli. Then, by measuring the volume of carbon monoxide absorbed in a period of time and dividing this by the alveolar carbon monoxide partial pressure, one can determine accurately the carbon monoxide-diffusing capacity.

To convert carbon monoxide-diffusing capacity to oxygen-diffusing capacity, the value is multiplied by a factor of 1.23 because the diffusion coefficient for oxygen is 1.23 times that for carbon monoxide. Thus, the average diffusing capacity for carbon monoxide in young male adults is 17 ml per minute per mm Hg, and the diffusing capacity for oxygen is 1.23 times this, or 21 ml per minute per mm Hg.

EFFECT OF THE VENTILATION-PERFUSION RATIO ON ALVEOLAR GAS CONCENTRATION

In the early part of this chapter we learned that two factors determine the P_{O_2} and the P_{CO_2} in the alveoli: (1) the rate of alveolar ventilation, and (2) the rate of transfer of oxygen and carbon dioxide through the respiratory membrane. However, these earlier discussions made the assumption that all the alveoli are ventilated equally and also that the blood flow through the alveolar capillaries is the same for each alveolus. However, as we have learned more about pulmonary function, it has become clear that in many lung diseases some areas of the lungs are well ventilated but have almost no blood flow, whereas other areas may have excellent blood flow but little or no ventilation. Obviously, in either of these conditions, gas exchange through the respiratory membrane is seriously impaired, and the person may suffer severe respiratory distress despite both normal *total* ventilation and also normal *total* pulmonary blood flow, but with the ventilation and blood flow going to different parts of the lungs. Therefore, a new concept has been developed to help us understand respiratory exchange when there is imbalance between alveolar ventilation and alveolar blood flow. This concept is called the *ventilation-perfusion ratio*.

In quantitative terms, the ventilation-perfusion ratio is expressed as $\dot{V}_A/\dot{Q}$. When $\dot{V}_A$ (alveolar ventilation) is normal for a given alveolus and $\dot{Q}$ (blood flow) is also normal for the same alveolus, then the ventilation-perfusion ratio ($\dot{V}_A/\dot{Q}$) is also said to be normal. However, when the ventilation ($\dot{V}_A$) is zero and yet there is still perfusion ($\dot{Q}$) of the alveolus, the ventilation-perfusion ratio is zero. Or, at the other extreme, when there is adequate ventilation ($\dot{V}_A$) but zero perfusion ($\dot{Q}$), then the ratio is infinity. At a ratio of either zero or infinity, there will be no exchange of gases through the respiratory membrane of the affected alveoli, which explains the importance of this concept. Therefore, let us explain the respiratory consequences of these two extremes and then consider the consequences of ratios between the extremes.

Alveolar Function When $\dot{V}_A/\dot{Q}$ Equals Zero. A $\dot{V}_A/\dot{Q}$ equal to zero means that there is zero alveolar ventilation but still blood flow through the alveolar capillaries. Obviously, without ventilation there will be zero gas exchange between the atmosphere and the blood for the simple reason that no gas can move between the alveoli and the exterior air.

However, even when $\dot{V}_A/\dot{Q}$ is equal to zero, the air in the alveolus still comes to equilibrium with the blood gases, because gases continue to diffuse between the blood and the alveolar air. Since the blood that perfuses the capillaries is venous blood returning to the lungs from the systemic circulation, it is the gases in this blood that come to equilibrium with the alveolar gases. In the next chapter we shall see that the normal venous blood ($\bar{v}$) has a P_{O_2} of 40 mm Hg and a P_{CO_2} of 45 mm Hg. Therefore, these are also the usual partial pressures of these two gases in alveoli that have blood flow but no ventilation.

Alveolar Function When $\dot{V}_A/\dot{Q}$ Equals Infinity. A $\dot{V}_A/\dot{Q}$ equal to infinity means that there is alveolar ventilation but no blood flow through the alveolar capillaries. Obviously, this condition is also incompatible with exchange of any gas whatsoever, for the simple reason that there is no blood to carry the gases away from the alveoli.

Yet, the effect on the alveolar gas partial pressures is entirely different from the effect when $\dot{V}_A/\dot{Q}$ equals zero. Instead of the alveolar gases coming to equilibrium with the venous blood, the alveolar air now becomes equal to the humidified inspired air. That is, the air that is inspired loses no oxygen to the blood and it gains no CO_2. And, since normal inspired and humidified air has a P_{O_2} of 149 mm Hg and a P_{CO_2} of 0 mm Hg, these are the partial pressures of these two gases in the alveoli.

Gas Exchange, and Alveolar Partial Pressures When $\dot{V}_A/\dot{Q}$ Is Normal. When there is both normal alveolar ventilation and also normal alveolar capillary blood flow (which means normal alveolar perfusion), exchange of oxygen and carbon dioxide through the respiratory membrane is approximately optimal. Oxygen is absorbed out of the inspired air so that the alveolar P_{O_2} lies somewhere between that of the inspired air and that of the venous blood. Likewise, carbon dioxide is transferred from the venous blood into the alveoli, which makes the alveolar P_{CO_2} rise to some level between that of the inspired air and that of the venous blood. Thus, under normal conditions, the alveolar air P_{O_2} averages 104 mm Hg and the P_{CO_2}, 40 mm Hg.

The P_{O_2}-P_{CO_2}, $\dot{V}_A/\dot{Q}$ Diagram. The concepts presented in the above sections can also be shown very usefully in graphical form, as illustrated in Figure 40–11, in a diagram called the "P_{O_2}-P_{CO_2}, $\dot{V}_A/\dot{Q}$ diagram." The curve in the diagram represents all possible P_{O_2} and P_{CO_2} combinations between the limits of $\dot{V}_A/\dot{Q}$ equals zero and $\dot{V}_A/\dot{Q}$ equals infinity when the gaseous pressures in the venous blood are normal and the person is breathing air at sea level pressure. Thus, point $\bar{v}$ is the plot of P_{O_2} and P_{CO_2} when $\dot{V}_A/\dot{Q}$ equals zero. At this point, the P_{O_2} is 40 mm Hg and the P_{CO_2} is 45 mm Hg, which are the values in venous blood.

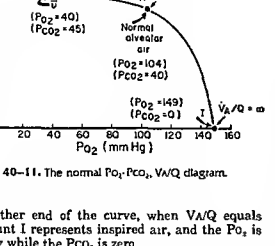


Figure 40–11. The normal P_{O_2}-P_{CO_2}, VA/Q diagram.

At the other end of the curve, when VA/Q equals infinity, point I represents inspired air, and the P_{O_2} is 149 mm Hg while the P_{CO_2} is zero

Also plotted on the curve is the point representing normal alveolar air when the VA/Q is normal. At this point, P_{O_2} is 104 mm Hg and P_{CO_2} is 40 mm Hg.

However, note once again that as VA/Q decreases below normal—that is, alveolar ventilation becomes too little in relation to alveolar perfusion—the alveolar P_{O_2} and P_{CO_2} approach those values found in venous blood. On the other hand, as VA/Q rises above normal and approaches infinity, the alveolar air composition approaches that of inspired air.

The Concept of "Physiologic Shunt" (When VA/Q Is Below Normal)

Whenever VA/Q is below normal, there obviously is not ventilation enough to provide the oxygen needed to oxygenate the blood flowing through the alveolar capillaries. Therefore, a certain *fraction of the venous blood* passing through the pulmonary capillaries does not become oxygenated. This fraction is called *shunted blood*. Also, some additional blood flows through the bronchial vessels rather than through the alveolar capillaries, normally about 2 per cent of the cardiac output; this too is unoxygenated, shunted blood.

The total quantitative amount of shunted blood per minute is called the *physiologic shunt*. This physiologic shunt is measured in the clinical pulmonary function laboratory by analyzing the concentration of oxygen in both mixed venous blood and arterial blood along with simultaneous measurement of the cardiac output. From these values the physiologic shunt can then be calculated using the following equation:

$$\frac{Q_{PS}}{Q_T} = \frac{C_{a_{O_2}} - C_{a_{O_2}}}{C_{a_{O_2}} - C_{\bar{v}_{O_2}}}$$

in which:

Q_{PS} is the physiologic shunt blood flow per minute

Q_T is cardiac output per minute

C_a is the concentration of oxygen in the arterial blood if there is "ideal" ventilation-perfusion ratio

$C_{a_{O_2}}$ is the measured concentration of oxygen in the arterial blood

$C_{\bar{v}_{O_2}}$ is the measured concentration of oxygen in the mixed venous blood

Obviously, the greater the physiologic shunt, the greater the amount of blood that fails to be oxygenated as it passes through the lungs.

The Concept of "Physiologic Dead Space" (When VA/Q Is Greater Than Normal)

When ventilation of some of the alveoli is great but the alveolar blood flow is low, there is then far more available oxygen in the alveoli than can be transported away from the alveoli by the flowing blood. Thus, the ventilation of these alveoli is said to be *wasted*. And, in addition, the ventilation of the anatomical dead space areas of the respiratory passageways is also wasted. The *sum of these two types of wasted ventilation is called the physiologic dead space*. This is measured in the clinical pulmonary function laboratory by making appropriate blood and expiratory gas measurements and using the following equation, called the Bohr equation:

$$\frac{V_{D_{phys}}}{V_T} = \frac{P_{a_{CO_2}} - P\bar{E}_{CO_2}}{P_{a_{CO_2}}}$$

in which

$V_{D_{phys}}$ is the physiologic dead space

V_T is the tidal volume

$P_{a_{CO_2}}$ is the partial pressure of carbon dioxide in the arterial blood

$P\bar{E}_{CO_2}$ is the average partial pressure of carbon dioxide in the entire expired air

Obviously, when the physiologic dead space is very great, much of the work of ventilation is wasted effort because so much of the ventilated air never reaches the blood.

Abnormalities of Ventilation-to-Perfusion Ratio

Abnormal VA/Qs in the Upper and Lower Normal Lung. In a normal person in the upright position, both blood flow and alveolar ventilation are considerably less in the upper part of the lung than in the lower part; however, blood flow is decreased considerably more than ventilation. Therefore, at the top of the lung, VA/Q is as much as three times as great as the ideal value, which causes a moderate degree of *physiologic dead space* in this area of the lung.

At the other extreme, in the bottom of the lung there is slightly too little ventilation in relation to blood flow, with VA/Q as low as 0.6 times the ideal value. Therefore, in this area a small fraction of the blood fails to become normally oxygenated, and this represents a *physiologic shunt*.

Therefore, in both extremes of the lung, inequalities of ventilation and perfusion decrease the effectiveness of the lung slightly for exchange of oxygen and carbon dioxide. However, during exercise the blood flow to the upper part of the lung increases markedly so that then far less physiologic dead space occurs and the effectiveness of gas exchange approaches optimum.

Abnormal VA/Q in Chronic Obstructive Lung Disease. Most persons who smoke for prolonged periods of time develop some degree of bronchial obstruction; indeed, in a large share of them this eventually becomes

104 mm Hg, whereas the Po_2 of the venous blood entering the capillary is only 40 mm Hg because a large amount of oxygen has been removed from this blood as it has passed through the body. Therefore, the pressure difference that causes oxygen to diffuse into the pulmonary capillary is 104 − 40 or 64 mm Hg. The curve below the capillary shows the progressive rise in blood Po_2 as the blood passes through the capillary, showing that the Po_2 rises essentially to equal that of the alveolar air by the time the blood has moved a third of the distance through the capillary, becoming approximately 104 mm Hg.

Uptake of Oxygen by the Pulmonary Blood During Exercise. During strenuous exercise, a person's body may require as much as 20 times the normal amount of oxygen. Also, because of the increased cardiac output, the time that the blood remains in the capillary may be reduced to less than one half normal despite the fact that additional capillaries open up. Therefore, oxygenation of the blood could suffer for two reasons. Yet, because of the great *safety factor for diffusion* of oxygen through the pulmonary membrane, the blood is still *almost completely saturated* with oxygen when it leaves the pulmonary capillaries. The reasons for this are:

First, it was pointed out in the previous chapter that the diffusing capacity for oxygen increases about threefold during exercise; this results mainly from increased numbers of capillaries participating in the diffusion but also from dilatation of both the alveoli and capillaries, as well as a more ideal ventilation-perfusion ratio in the upper part of the lungs.

Second, note in Figure 41–1 that during normal pulmonary blood flow the blood becomes almost saturated with oxygen by the time it has passed through one third of the pulmonary capillary, and little additional oxygen enters the blood during the latter two thirds of its transit. That is, the blood normally stays in the lung capillaries about three times as long as necessary to cause full oxygenation. Therefore, even with the shortened time of exposure in exercise, the blood still can become either fully oxygenated or nearly so.

TRANSPORT OF OXYGEN IN THE ARTERIAL BLOOD

About 98 per cent of the blood that enters the left atrium from the lung passes through the alveolar capillaries and becomes fully oxygenated, that is, up to a Po_2 of approximately 104 mm Hg as just explained. However, another 2 per cent of the blood leaving the lungs passes through the [illegible]al circulation, which supplies mainly the [illegible] tissues of the lungs and is not exposed to the [illegible] air. Therefore, this blood flow represents "shunt" flow, meaning blood that is shunted past the gas exchange areas. On leaving the lungs,

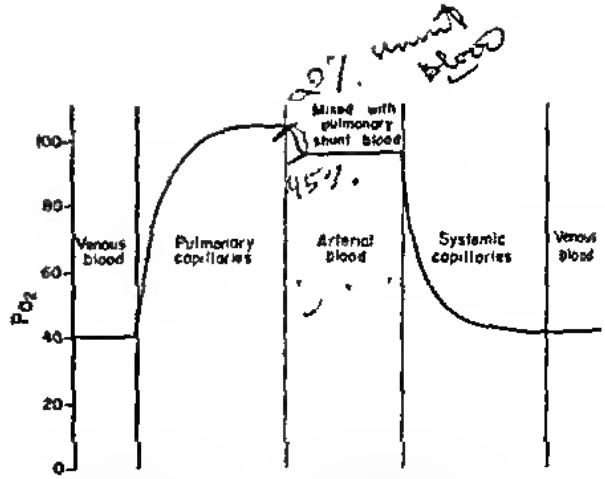


Figure 41–2. Changes in Po_2 in the pulmonary capillary blood, the arterial blood, and the systemic capillary blood, illustrating the effect of "venous admixture."

its Po_2 is approximately that of normal venous blood, about 40 mm Hg. This blood combines in the pulmonary veins with the oxygenated blood from the alveolar capillaries; this mixing of the bloods is called *venous admixture of blood*, and it causes the Po_2 of the blood pumped by the left heart into the aorta to fall to approximately 95 mm Hg. These changes in blood Po_2 at different points in the circulatory system are illustrated in Figure 41–2.

DIFFUSION OF OXYGEN FROM THE CAPILLARIES TO THE INTERSTITIAL FLUID

When the arterial blood reaches the peripheral tissues, its Po_2 is still 95 mm Hg. On the other hand, as shown in Figure 41–3, the Po_2 in the interstitial fluid averages only 40 mm Hg. Thus, there is a tremendous pressure difference that causes oxygen to diffuse very rapidly from the blood into the tissues, so rapidly that the capillary Po_2 falls almost to equal the 40 mm Hg pressure in the interstitium. Therefore, the Po_2 of the blood entering the veins from the tissue capillaries is also about 40 mm Hg.

Effect of Rate of Blood Flow on Interstitial Fluid Po_2. If the blood flow through a particular tissue becomes increased, greater quantities of oxygen are transported into the tissue in a given period of time, and the tissue Po_2 becomes correspondingly increased. This effect is illustrated in Figure 41–4. Note that an increase in flow to 400 per cent of normal increases the Po_2 from 40 mm Hg at Point A in the figure to 66 mm Hg at Point B. However, the upper limit to which the Po_2 can

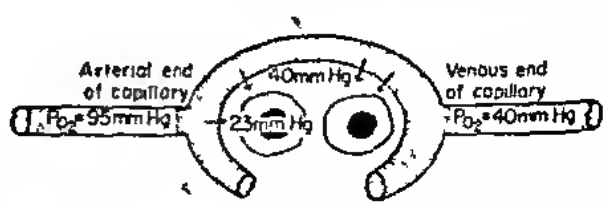


Figure 41–3. Diffusion of oxygen from a tissue capillary to the cells.

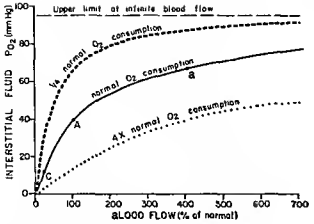


Figure 41–4. Effect of blood flow and rate of oxygen consumption on tissue Po_2

rise, even with maximal blood flow, is 95 mm Hg, because this is the oxygen pressure in the arterial blood.

Effect of Rate of Tissue Metabolism on Interstitial Fluid Po_2. If the cells utilize more oxygen for metabolism than normally, this tends to reduce the interstitial fluid Po_2. Figure 41–4 also illustrates this effect, showing reduced interstitial fluid Po_2 when the oxygen consumption is increased, and increased Po_2 when the consumption is decreased

Effect of Hemoglobin Concentration on Interstitial Fluid Po_2. Because approximately 97 per cent of the oxygen transported in the blood is carried by hemoglobin, a decrease in hemoglobin concentration has the same effect on interstitial fluid Po_2 as a decrease in blood flow. Thus, reducing the hemoglobin concentration to one quarter normal while maintaining normal blood flow reduces the interstitial fluid Po_2 to the same level as point C in Figure 41–4—that is, to about 13 mm Hg.

In summary, tissue Po_2 is determined by a balance between (a) the rate of oxygen transport to the tissues in the blood and (b) the rate at which the oxygen is utilized by the tissues.

DIFFUSION OF OXYGEN FROM THE CAPILLARIES TO THE CELLS

Since oxygen is always being used by the cells, the intracellular Po_2 remains lower than the interstitial fluid Po_2 that immediately surrounds the capillaries. Also, in many instances, there is considerable distance between the capillaries and the cells. Therefore, the normal intracellular Po_2 ranges from as low as 5 mm Hg to as high as 40 mm Hg, averaging (by direct measurement in lower animals) 23 mm Hg. Since only 3 mm Hg of oxygen pressure is normally required for full support of the metabolic processes of the cell, one can see that even this low cellular Po_2, 23 mm Hg, is more than adequate and actually provides a considerable safety factor.

DIFFUSION OF CARBON DIOXIDE FROM THE CELLS TO THE TISSUE CAPILLARIES, AND FROM THE PULMONARY CAPILLARIES TO THE ALVEOLI

When oxygen is used by the cells, most of it becomes carbon dioxide, and this increases the

Figure 41–5. Uptake of carbon dioxide by the blood in the capillaries

intracellular Pco_2. Therefore, carbon dioxide diffuses from the cells into the tissue capillaries and is then carried by the blood to the lungs, where it diffuses from the pulmonary capillaries into the alveoli. Thus, at each point in the gas transport chain, carbon dioxide diffuses in exactly the opposite direction to the diffusion of oxygen. Yet, there is one major difference between the diffusion of carbon dioxide and that of oxygen: carbon dioxide diffuses about 20 times as rapidly as oxygen. Therefore, the pressure differences that cause carbon dioxide diffusion are, in each instance, far less than the pressure differences required to cause oxygen diffusion. These pressures are the following:

(1) Intracellular Pco_2, about 46 mm Hg; interstitial Pco_2, about 45 mm Hg; thus, there is only a 1 mm Hg pressure differential, as illustrated in Figure 41–5.

(2) Pco_2 of the arterial blood entering the tissues, 40 mm Hg; Pco_2 of the venous blood leaving the tissues, about 45 mm Hg; thus, as also illustrated in Figure 41–5, the tissue capillary blood comes almost exactly to equilibrium with the interstitial Pco_2, also 45 mm Hg.

(3) Pco_2 of the venous blood entering the pulmonary capillaries, 45 mm Hg; Pco_2 of the alveolar air, 40 mm Hg; thus, only a 5 mm Hg pressure difference causes all the required carbon dioxide diffusion out of the pulmonary capillaries into the alveoli. Furthermore, as illustrated in Figure 41–6, the Pco_2 of the pulmonary capillary blood falls

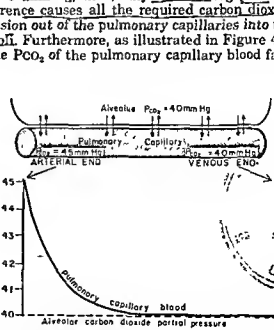


Figure 41–6. Diffusion of carbon dioxide from the pulmonary blood into the alveolus. (This curve was constructed from data in Milhorn and Pulley. *Biophys. J.*, 8:337, 1968.)

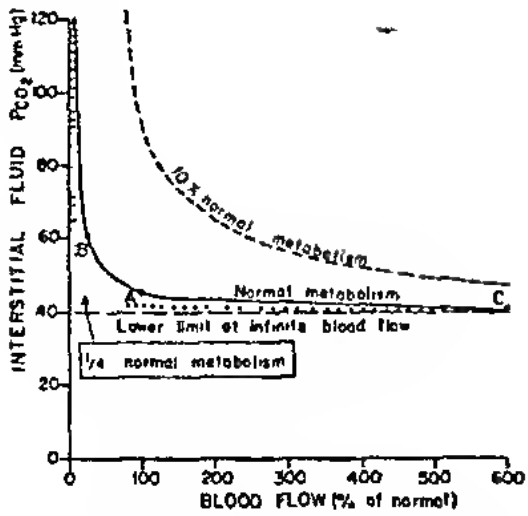


Figure 41–7. Effect of blood flow and metabolic rate on tissue P_{CO_2}

almost exactly to equal the alveolar P_{CO_2} of 40 mm Hg before it has passed more than about one third the distance through the capillaries. This is the same effect that was observed earlier for oxygen diffusion as well

Effect of Tissue Metabolism and Blood Flow on Interstitial P_{CO_2}. Tissue capillary blood flow and tissue metabolism affect the P_{CO_2} exactly opposite to the way that they affect tissue P_{O_2}. Figure 41–7 shows these effects, as follows:

(1) A decrease in blood flow from normal, point A, to one quarter normal, point B, increases the tissue P_{CO_2} from the normal value of 45 mm Hg up to the elevated level of 60 mm Hg Conversely, increasing the blood flow to six times normal, point C, decreases the P_{CO_2} from the normal value of 45 mm Hg to 41 mm Hg, almost down to a level equal to the P_{CO_2} in the arterial blood, 40 mm Hg, entering the tissue capillaries

(2) Note also that a tenfold increase in metabolic rate greatly elevates the P_{CO_2} at all levels of blood flow, whereas decreasing the metabolism to one quarter normal causes the interstitial fluid P_{CO_2} to fall to about 41 mm Hg, closely approaching that of the arterial blood, 40 mm Hg.

TRANSPORT OF OXYGEN IN THE BLOOD

Normally, about 97 per cent of the oxygen transported from the lungs to the tissues is carried in chemical combination with hemoglobin in the red blood cells, and the remaining 3 per cent is carried in the dissolved state in the water of the plasma and cells. Thus, *under normal conditions* oxygen is carried to the tissues almost entirely by hemoglobin.

THE REVERSIBLE COMBINATION OF OXYGEN WITH HEMOGLOBIN

The chemistry of hemoglobin was presented in Chapter 4, where it was pointed out that the oxygen molecule combines loosely and reversibly with the heme portion of the hemoglobin. When the P_{O_2} is high, as in the pulmonary capillaries oxygen binds with the hemoglobin, but when the P_{O_2} is low, as in the tissue capillaries, oxygen is released from the hemoglobin This is the basis for oxygen transport from the lungs to the tissues.

The Oxygen-Hemoglobin Dissociation Curve. Figure 41–8 illustrates the oxygen-hemoglobin dissociation curve, which shows a progressive increase in the per cent of the hemoglobin that is bound with oxygen as the P_{O_2} increases which is called the *per cent saturation of the hemoglobin*. Since the blood in the arteries usually has a P_{O_2} of about 95 mm Hg, one can see from the dissociation curve that the *usual oxygen saturation of arterial blood is about 97 per cent*. On the other hand, in normal venous blood returning from the tissues the P_{O_2} is about 40 mm Hg and *the saturation of the hemoglobin is about 75 per cent*.

Maximum Amount of Oxygen That Can Combine with the Hemoglobin of the Blood

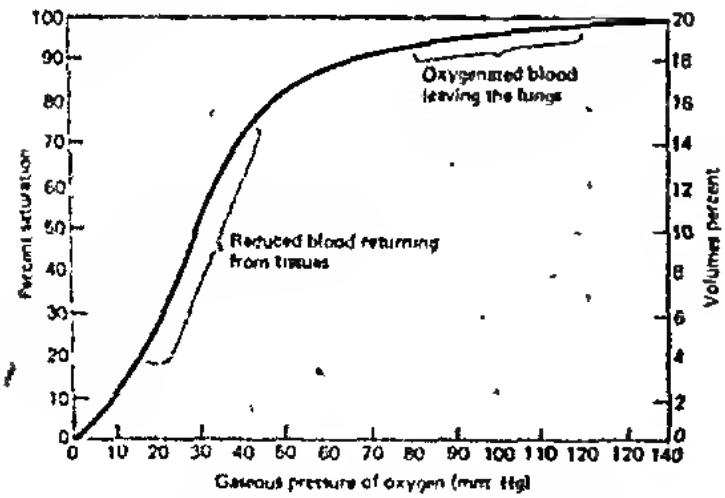


Figure 41–8. The oxygen hemoglobin dissociation curve

The blood of a normal person contains approximately 15 grams of hemoglobin in each 100 ml of blood, and each gram of hemoglobin can bind with a maximum of about 1.34 ml of oxygen (1.39 ml when the hemoglobin is chemically pure, but this is reduced about 4 per cent by impurities such as methemoglobin). Therefore, on the average, the hemoglobin in 100 ml of blood can combine with a total of almost exactly 20 ml of oxygen when the hemoglobin is 100 per cent saturated. This is usually expressed as 20 *volumes per cent*. The oxygen-hemoglobin dissociation curve for the normal person, therefore, can also be expressed in terms of volume per cent of oxygen, as shown by the right-most scale in Figure 41–8, as well as in per cent saturation of hemoglobin.

Amount of Oxygen Released from the Hemoglobin in the Tissues. The total quantity of oxygen *bound with hemoglobin* in normal arterial blood, which is 97 per cent saturated, is approximately 19.4 ml per 100 ml of blood. This is illustrated in Figure 41–9. However, on passing through the tissue capillaries, this amount is reduced, on the average, to 14.4 ml (Po_2 of 40 mm Hg, 75 per cent saturated). Thus, *under normal conditions about 5 ml of oxygen is transported to the tissues by each 100 ml of blood.*

Transport of Oxygen During Strenuous Exercise. In heavy exercise the muscle cells utilize oxygen at a rapid rate, which causes the interstitial fluid Po_2 to fall to as low as 15 mm Hg. At this pressure only 4.4 ml of oxygen remains bound with the hemoglobin in each 100 ml of blood, as also shown in Figure 41–9. Thus 19.4 − 4.4, or 15 ml, is the quantity of oxygen then transported by each 100 ml of blood. Thus, three times as much oxygen is transported in each volume of blood that passes through the tissues as normally. And when one remembers that the cardiac output can also increase to six to seven times normal in well-trained marathon runners, multiplying these two gives a 20-fold increase in oxygen transport to the tissues: this is about the limit that can be achieved.

The Utilization Coefficient. The percentage of the blood that gives up its oxygen as it passes through the tissue capillaries is called the *utilization coefficient*. The normal value for this is approximately 25 per cent, as was evident from the preceding discussion. However, during strenuous exercise the utilization coefficient can increase to as much as 75 to 85 per cent. And, in local tissue areas where the blood flow is very slow or the metabolic rate very high, utilization coefficients approaching 100 per cent have been recorded—that is, essentially all the oxygen is removed.

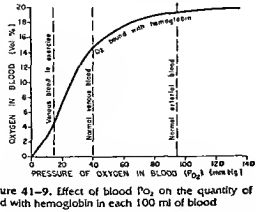


Figure 41–9. Effect of blood Po_2 on the quantity of oxygen bound with hemoglobin in each 100 ml of blood

THE TISSUE OXYGEN BUFFER FUNCTION OF HEMOGLOBIN

Though hemoglobin is necessary for transport of oxygen to the tissues, it performs still another major function essential to life. This is its function as a "tissue oxygen buffer" system. That is, the hemoglobin in the blood is mainly responsible for controlling the oxygen pressure in the tissues. This can be explained as follows:

Role of Hemoglobin in Maintaining Constant Po_2 in the Tissues. Under basal conditions the tissues require about 5 ml of oxygen from each 100 ml of blood passing through the tissue capillaries. Referring back to the oxygen-hemoglobin dissociation curve in Figure 41–9, one can see that for the 5 ml of oxygen to be released, the Po_2 must fall to about 40 mm Hg. Therefore, the tissue Po_2 normally cannot rise above this 40 mm Hg level, for if such should occur, the oxygen needed by the tissues could not be released from the hemoglobin. In this way, the hemoglobin normally sets an upper limit on the gaseous pressure in the tissues at approximately 40 mm Hg.

On the other hand, in heavy exercise extra large amounts of oxygen must be delivered from the hemoglobin to the tissues. But this can be achieved with very little further decrease in tissue Po_2 because of the steep slope of the dissociation curve—that is, a small fall in Po_2 causes extreme amounts of oxygen to be released. Therefore, the Po_2 only rarely falls below 20 mm Hg.

It can be seen, then, that the hemoglobin in the blood automatically delivers oxygen to the tissues at a pressure between approximately 20 and 40 mm Hg. This seems to be a wide range of Po_2s in the interstitial fluid, but, when one considers how much the interstitial fluid Po_2 might possibly change during exercise and other types of stress, this range of 20 to 40 mm Hg is relatively narrow.

Value of Hemoglobin for Maintaining Constant Tissue Po_2 When Atmospheric Oxygen Concentration Changes Markedly. The normal Po_2 in the alveoli is approximately 104 mm Hg, but, as one ascends a mountain or goes high in an airplane, the Po_2 can easily fall to less than half this. Or, when one enters areas of compressed air, such as deep below the sea or in pressurized chambers, the Po_2 may rise to ten times this level. Yet, even so, the tissue Po_2 changes very little. Let us explain this.

It will be seen from the oxygen-hemoglobin dissociation curve in Figure 41–8 that when the alveolar PO_2 is decreased to as low as 60mm Hg the arterial hemoglobin is still 89 per cent saturated, only 8 per cent below the normal saturation of 97 per cent. Furthermore, the tissues still remove approximately 5 ml of oxygen from each 100 ml of blood passing through the tissues; to remove this oxygen, the PO_2 of the venous blood falls to 35 mm Hg, only 5 mm below the normal value. Thus, the tissue PO_2 hardly changes despite the marked fall in alveolar PO_2 from 104 to 60 mm Hg.

On the other hand, when the alveolar PO_2 rises far above the normal value of 104 mm Hg, the maximum oxygen saturation of hemoglobin can never rise above 100 per cent. Therefore, even though the oxygen in the alveoli should rise to a partial pressure of 500 mm Hg, or even more, the saturation of hemoglobin would increase only 3 per cent because, even at 104 mm Hg PO_2, 97 per cent of the hemoglobin is already combined with oxygen; and only a small amount of additional oxygen dissolves in the fluid of the blood, as will be discussed subsequently. Then, when the blood passes through the tissue capillaries, it still loses several milliliters of oxygen to the tissues, and this loss automatically reduces the PO_2 of the capillary blood to a value only a few millimeters greater than the normal 40 mm Hg.

Consequently, the level of alveolar oxygen may vary greatly—from 60 to more than 500 mm Hg PO_2—and still the PO_2 in the tissue does not vary more than a few millimeters from normal, illustrating beautifully the tissue oxygen buffer function of the blood hemoglobin.

SHIFT OF THE OXYGEN-HEMOGLOBIN DISSOCIATION CURVE AND ITS SIGNIFICANCE

The oxygen-hemoglobin dissociation curves of Figures 41–8 and 41–9 are those for normal average blood. However, a number of different factors can displace the dissociation curve in one direction or the other in the manner illustrated in Figure 41–10. This figure shows that when the blood becomes slightly acidic, with the pH decreasing from the normal value of 7.4 to 7.2, the oxygen-hemoglobin dissociation curve shifts on the average about 15 per cent to the right. On the other hand, an increase in the pH to 7.6 shifts the curve a similar amount to the left.

In addition to pH changes, several other factors are also known to shift the curve. Three of these, all of which shift the curve to the right, are (1) increased carbon dioxide concentration, (2) increased blood temperature, and (3) increased 2,3-diphosphoglycerate, a phosphate compound normally present in the blood but in differing concentrations under different conditions.

Another condition that shifts the dissociation curve to the left is the presence in the blood of large quantities of *fetal hemoglobin*, a type of hemoglobin normally present in the fetus before birth and which is different from the normal hemoglobin called *adult hemoglobin*. Shift of the curve to the left when fetal hemoglobin is

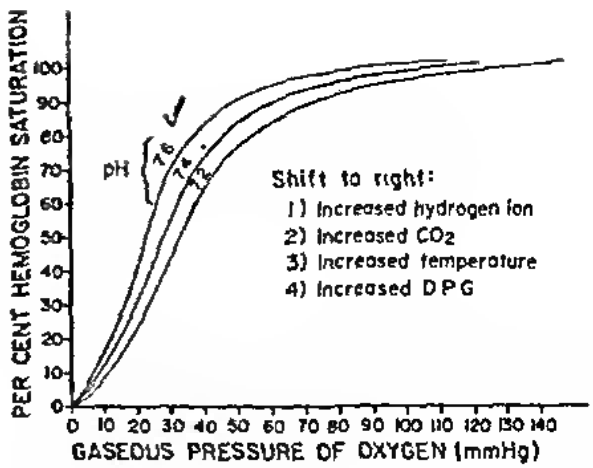


Figure 41–10. Shift of the oxygen-hemoglobin dissociation curve to the right by increases in (1) hydrogen ions, (2) CO_2, (3) temperature, or (4) DPG

present is important for oxygen delivery to the fetal tissues under the hypoxic conditions in which the fetus exists. This will be discussed further in Chapter 82.

Importance of the Effect of Carbon Dioxide and Hydrogen Ions in Shifting the Oxygen-Hemoglobin Dissociation Curve—the Bohr Effect. Shift of the oxygen-hemoglobin dissociation curve by changes in the blood CO_2 and hydrogen ions is important to enhance oxygenation of the blood in the lungs and also to enhance release of oxygen from the blood in the tissues. This is called the *Bohr effect*, and it can be explained as follows. As the blood passes through the lungs, carbon dioxide diffuses from the blood into the alveoli. This reduces the blood P_{CO_2} and also decreases the hydrogen ion concentration because of the resulting decrease in blood carbonic acid. Both these effects shift the oxygen-hemoglobin dissociation curve to the left and upward. Therefore, the quantity of oxygen that binds with the hemoglobin at any given alveolar PO_2 becomes considerably increased, thus allowing greater oxygen transport to the tissues. Then when the blood reaches the tissue capillaries, exactly the opposite effect occurs. Carbon dioxide entering the blood from the tissues shifts the curve rightward, which displaces oxygen from the hemoglobin and therefore delivers oxygen to the tissues at a higher PO_2 than would otherwise occur.

Effect of 2,3-Diphosphoglycerate (DPG). The normal DPG in the blood keeps the oxygen-hemoglobin dissociation curve shifted slightly to the right all the time. In addition, in hypoxic conditions that last longer than a few hours, the quantity of DPG in the blood increases considerably, thus shifting the oxygen-hemoglobin dissociation curve even farther to the right. This causes oxygen to be released to the tissues at as much as 10 mm Hg higher oxygen pressure than would be the case without this increased DPG. Therefore, it has been taught for the past few years that this might be an important mechanism for adaptation to hypoxia. However, the presence of the excess DPG also makes it difficult for the hemoglobin to combine with oxygen in the lungs when the alveolar PO_2 is reduced, thereby often creating as much harm as good. Therefore, it is questionable whether the DPG shift of the dissociation curve is as beneficial in hypoxia as had been believed.

Shift of the Dissociation Curve During Exercise. In exercise, several factors shift the dissociation curve

considerably to the right. The exercising muscles release large quantities of carbon dioxide and this, plus acids released by the exercising muscle, increases the hydrogen ion concentration in the muscle capillary blood. In addition, the temperature of the muscle often rises as much as 3 to 4 degrees, and, finally, phosphate compounds are also released. All these factors acting together shift the oxygen-hemoglobin dissociation curve *of the muscle capillary blood* considerably to the right. Therefore, oxygen can sometimes be released to the muscle at Po_2s as great as 40 mm Hg (the normal resting value) even though as much as 75 to 85 per cent of the oxygen is removed from the hemoglobin. Then, in the lungs, the shift occurs in the opposite direction, thus allowing pickup of extra amounts of oxygen from the alveoli.

METABOLIC USE OF OXYGEN BY THE CELLS

Relationship Between Intracellular Po_2 and Rate of Oxygen Usage. Only a minute level of oxygen pressure is required in the cells for normal intracellular chemical reactions to take place. The reason for this is that the respiratory enzyme systems of the cell, which will be discussed in Chapter 67, are geared so that when the cellular Po_2 is more than 1 to 3 mm Hg, oxygen availability is no longer a limiting factor in the rates of the chemical reactions. Instead, the main limiting factor then is the *concentration of adenosine diphosphate* (ADP) in the cells, as was explained in Chapter 3. This effect is illustrated in Figure 41–11, which shows the relationship between intracellular Po_2 and rate of oxygen usage. Note that whenever the intracellular Po_2 is above about 3 mm Hg the rate of oxygen usage becomes constant for any given concentration of ADP in the cell. On the other hand, when the ADP concentration is altered, the rate of oxygen usage changes in proportion to the change in ADP concentration.

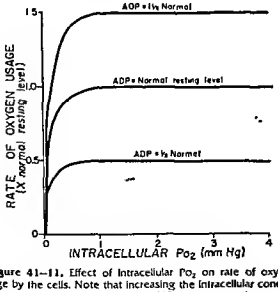


Figure 41–11. Effect of intracellular Po_2 on rate of oxygen usage by the cells. Note that increasing the intracellular concentration of adenosine diphosphate (ADP) increases the rate of oxygen usage.

It will be recalled from the discussion in Chapter 3 that when adenosine triphosphate (ATP) is utilized in the cells to provide energy, it is converted into ADP. The increasing concentration of ADP, in turn, increases the metabolic usage of both oxygen and the various nutrients that combine with the oxygen to release energy. This energy is used to reform the ATP. Therefore, *under normal operating conditions the rate of oxygen utilization by the cells is controlled ultimately by the rate of energy expenditure within the cells—that is, by the rate at which ADP is formed from ATP.* Only in very hypoxic states does the availability of oxygen become a limiting condition.

Effect of Diffusion Distance from the Capillary to the Cell on Oxygen Utilization. Cells are rarely more than 50 microns away from a capillary, and oxygen normally can diffuse readily enough from the capillary to the cell to supply all the required amounts of oxygen for metabolism. However, occasionally cells are located farther than this from the capillaries, and the rate of oxygen diffusion to these cells is so low that the intracellular Po_2 falls then below the critical level of 1 to 3 mm Hg required to maintain maximum intracellular metabolism. Thus, under these conditions, oxygen utilization by the cells is said to be *diffusion limited* and is no longer determined by the amount of ADP formed in the cells.

Effect of Blood Flow on Metabolic Use of Oxygen. The total amount of oxygen available each minute for use in any given tissue is determined by (1) the quantity of oxygen transported in each 100 ml of blood and (2) the rate of blood flow. If the rate of blood flow falls to zero, the amount of available oxygen obviously also falls to zero. Thus, there are times when the rate of blood flow through a tissue can be so low that the tissue Po_2 falls below the critical 1 to 3 mm Hg required for maximal intracellular metabolism. Under these conditions, the rate of tissue utilization of oxygen is *blood flow limited.* However, neither diffusion-limited nor blood flow–limited usage of oxygen can continue for long, because the cells then often receive less oxygen than is required for continuation of the life of the cells themselves.

TRANSPORT OF OXYGEN IN THE DISSOLVED STATE

At the normal arterial Po_2 of 95 mm Hg, approximately 0.29 ml of oxygen is dissolved in every 100 ml of water in the blood. Then when the Po_2 of the blood falls to 40 mm Hg in the tissue capillaries, only 0.12 ml of oxygen remains dissolved. In other words, 0.17 ml of oxygen is normally transported in the dissolved state to the tissues by each 100 ml of blood water. This compares with about 5.0 ml transported by the hemoglobin. Therefore, the amount of oxygen transported to the tissues in the dissolved state is normally slight, only about 3 per cent of the total as compared with 97 per cent transported by the hemoglobin. During strenuous exercise, when transport increases another threefold, the relative

quantity then transported in the dissolved state falls to as little as 1 5 per cent. Yet, if a person breathes oxygen at very high alveolar P_{O_2}s, the amount then transported in the dissolved state can become tremendous, so much so that serious excesses of oxygen occur in the tissues and "oxygen poisoning" ensues This often leads to convulsions and even death, as will be discussed in detail in Chapter 45 in relation to high pressure breathing

COMBINATION OF HEMOGLOBIN WITH CARBON MONOXIDE

Carbon monoxide combines with hemoglobin at the same point on the hemoglobin molecule as does oxygen Furthermore, it binds with about 230 times as much tenacity as oxygen, which is illustrated by the carbon monoxide-hemoglobin dissociation curve in Figure 41–12 This curve is almost identical with the oxygen-hemoglobin dissociation curve, except that the pressures of the carbon monoxide shown on the abscissa are at a level $1/230$ of those in the oxygen-hemoglobin dissociation curve of Figure 41–8 Therefore, a carbon monoxide pressure of only 0 4 mm Hg in the alveoli, $1/230$ that of the alveolar oxygen, allows the carbon monoxide to compete equally with the oxygen for combination with the hemoglobin and causes half the hemoglobin in the blood to become bound with carbon monoxide instead of with oxygen A carbon monoxide pressure of 0 7 mm Hg (a concentration of about 0 1 per cent in the air) can be lethal

A patient severely poisoned with carbon monoxide can be advantageously treated by administering pure oxygen, for oxygen at high alveolar pressures displaces carbon monoxide from its combination with hemoglobin far more rapidly than can oxygen at the low pressure of atmospheric oxygen

The patient can also be benefited by simultaneous administration of a few per cent carbon dioxide because this strongly stimulates the respiratory center, as is discussed in the following chapter This increases alveolar ventilation and reduces the alveolar carbon monoxide concentration, which allows increased carbon monoxide release from the blood

With intensive oxygen and carbon dioxide therapy, carbon monoxide can be removed from the blood 10 to 20 times as rapidly as without therapy

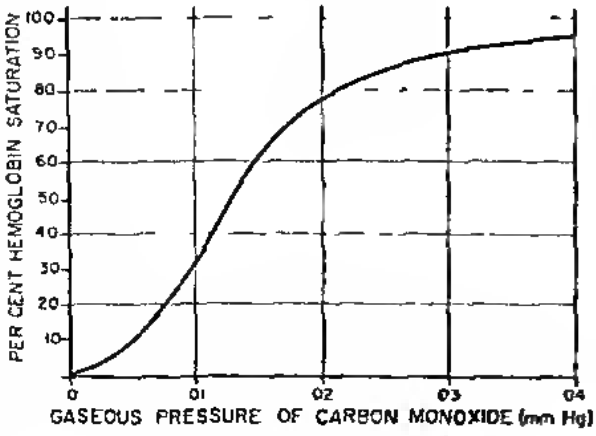


Figure 41–12. The carbon monoxide–hemoglobin dissociation curve

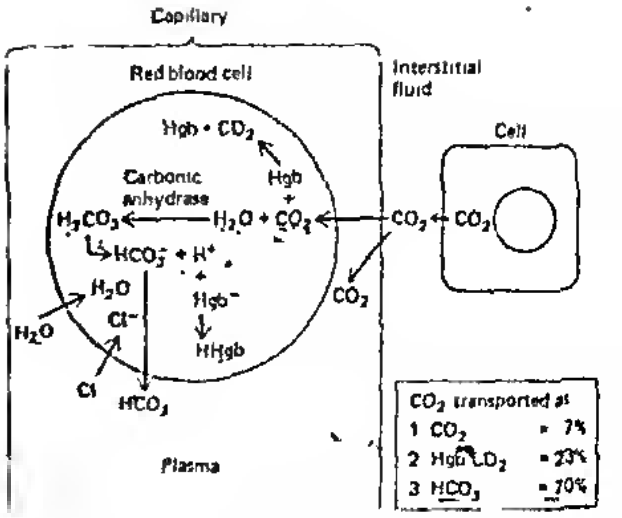


Figure 41–13. Transport of carbon dioxide in the blood

TRANSPORT OF CARBON DIOXIDE IN THE BLOOD

Transport of carbon dioxide by the blood is not nearly so great a problem as transport of oxygen, because even in the most abnormal conditions carbon dioxide can usually be transported in far greater quantities than can oxygen. However, the amount of carbon dioxide in the blood does have much to do with acid-base balance of the body fluids, which was discussed in Chapter 37.

Under normal resting conditions *an average of 4 ml of carbon dioxide is transported from the tissues to the lungs in each 100 ml of blood.*

CHEMICAL FORMS IN WHICH CARBON DIOXIDE IS TRANSPORTED

To begin the process of carbon dioxide transport, carbon dioxide diffuses out of the tissue cells in the gaseous form (but not to any significant extent in the bicarbonate form because the cell membrane is almost impermeable to bicarbonate ions) On entering the capillary, the carbon dioxide initiates a host of almost instantaneous physical and chemical reactions, illustrated in Figure 41–13, that are essential for CO_2 transport.

Transport of Carbon Dioxide in the Dissolved State. A small portion of the carbon dioxide is transported in the dissolved state to the lungs It will be recalled that the P_{CO_2} of venous blood is 45 mm Hg, and that of arterial blood is 40 mm Hg The amount of carbon dioxide dissolved in the fluid of the blood at 45 mm Hg is about 2.7 ml per dl (2.7 volumes per cent). The amount dissolved at 40 mm Hg is about 2 4 ml, or a difference of 0.3 ml. Therefore, only about 0 3 ml of carbon dioxide is transported in the form of dissolved carbon dioxide by each deciliter of blood. This is about 7 per cent of all the carbon dioxide transported.

Transport of Carbon Dioxide in the Form of Bicarbonate Ion. *Reaction of Carbon Dioxide with Water in the Red Blood Cells—Effect of Carbonic Anhydrase.* The dissolved carbon dioxide in the blood reacts with water to form carbonic acid. However, this reaction would occur much too slowly to be of importance were it not for the fact that inside the red blood cells the enzyme called *carbonic anhydrase* catalyzes the reaction between carbon dioxide and water, accelerating its rate about 5000-fold. Therefore, instead of requiring many seconds or minutes to occur, as is true in the plasma, the reaction occurs so rapidly in the red blood cells that it reaches almost complete equilibrium within a small fraction of a second. This allows tremendous amounts of carbon dioxide to react with the red cell water even before the blood leaves the tissue capillaries.

Dissociation of the Carbonic Acid into Bicarbonate and Hydrogen Ions. In another small fraction of a second the carbonic acid formed in the red cells dissociates into *hydrogen and bicarbonate ions.* Most of the hydrogen ions then combine with the hemoglobin in the red blood cells because hemoglobin is a powerful acid-base buffer. In turn, many of the bicarbonate ions diffuse into the plasma while chloride ions diffuse into the red cells to take their place. This is made possible by the presence of a special *bicarbonate-chloride carrier protein* in the red cell membrane that shuttles these two ions in opposite directions at rapid velocities. Thus, the chloride content of venous red blood cells is greater than that of arterial cells, a phenomenon called the *chloride shift.*

The reversible combination of carbon dioxide with water in the red blood cells under the influence of carbonic anhydrase probably accounts for at least 70 per cent of all the carbon dioxide transported from the tissues to the lungs. Thus, this means of transporting carbon dioxide is by far the most important of all the methods for transport. Indeed, when a carbonic anhydrase inhibitor (acetazolamide) is administered to an animal to block the action of carbonic anhydrase in the red blood cells, carbon dioxide transport from the tissues becomes very poor—so poor that the tissue PCO_2 can be made to rise to as high as 80 mm Hg instead of the normal 45 mm Hg.

Transport of Carbon Dioxide in Combination with Hemoglobin and Plasma Proteins—Carbaminohemoglobin. In addition to reacting with water, carbon dioxide also reacts directly with hemoglobin to form the compound *carbaminohemoglobin* ($Hgb{\cdot}CO_2$). This combination of carbon dioxide with the hemoglobin is a reversible reaction that occurs with a very loose bond, so that the carbon dioxide is easily released into the alveoli where the PCO_2 is lower than in the tissue capillaries. A small amount of carbon dioxide also reacts in this same way with the plasma proteins, but this is much less significant because the quantity of these proteins is only one fourth as great as the quantity of hemoglobin.

The *theoretical* quantity of carbon dioxide that can be carried from the tissues to the lungs in combination with hemoglobin and plasma proteins is approximately 30 per cent of the total quantity transported—that is, about 1.5 ml of carbon dioxide in each deciliter of blood. However, this reaction is much slower than the reaction of carbon dioxide with water inside the red blood cells. Therefore, it is doubtful that this mechanism provides transport of more than 15 to 25 per cent of the total quantity of carbon dioxide.

THE CARBON DIOXIDE DISSOCIATION CURVE

It is apparent that carbon dioxide can exist in the blood in many different forms: (1) as free carbon dioxide, and (2) in chemical combinations with water, hemoglobin, and plasma protein. The total quantity of carbon dioxide combined with the blood in all these forms depends on the PCO_2. The curve shown in Figure 41–14 depicts this dependence of total blood CO_2 in all its forms on PCO_2; this curve is called the *carbon dioxide dissociation curve.*

Note that the normal blood PCO_2 ranges between the limits of 40 mm Hg in arterial blood and 45 mm Hg in venous blood, which is a very narrow range. Note also that the normal concentration of carbon dioxide in the blood is about 50 volumes per cent but that only 4 volumes per cent of this is actually exchanged during normal transport of carbon dioxide from the tissues to the lungs. That is, the concentration rises to about 52 volumes per cent as the blood passes through the tissues, and falls to about 48 volumes per cent as it passes through the lungs.

EFFECT OF THE OXYGEN-HEMOGLOBIN REACTION ON CARBON DIOXIDE TRANSPORT—THE HALDANE EFFECT

Earlier in the chapter it was pointed out that an increase in carbon dioxide in the blood will cause oxygen

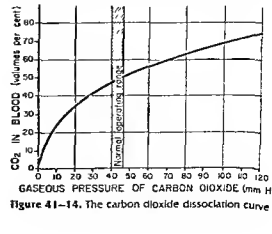


Figure 41–14. The carbon dioxide dissociation curve

to be displaced from the hemoglobin and that this is an important factor in promoting oxygen transport. The reverse is also true: binding of oxygen with hemoglobin tends to displace carbon dioxide from the blood. Indeed, this effect, called the *Haldane effect*, is quantitatively far more important in promoting carbon dioxide transport than is the Bohr effect in promoting oxygen transport.

The Haldane effect results from the simple fact that combination of oxygen with hemoglobin causes the hemoglobin to become a stronger acid. This in turn displaces carbon dioxide from the blood in two ways. (1) The more highly acidic hemoglobin has less tendency to combine with carbon dioxide to form carbaminohemoglobin, thus displacing much of the carbon dioxide that is present in this form in the blood. (2) The increased acidity of the hemoglobin causes it to release an excess of hydrogen ions, and these in turn bind with bicarbonate ions to form carbonic acid, then this dissociates into water and carbon dioxide, and the carbon dioxide is released from the blood into the alveoli. Thus, in the presence of oxygen, much less carbon dioxide can bind with the blood; and conversely, in the absence of oxygen, considerably more carbon dioxide can bind.

Therefore, in the tissue capillaries the Haldane effect causes increased pickup of carbon dioxide because of oxygen removal from the hemoglobin, and in the lungs it causes increased release of carbon dioxide because of oxygen pickup by the hemoglobin.

Figure 41–15 illustrates quantitatively the significance of the Haldane effect on the transport of carbon dioxide from the tissues to the lungs. This figure shows small portions of two separate carbon dioxide dissociation curves, the solid curve when the P_{O_2} is 100 mm Hg, which is the case in the lungs, and the dashed curve when the P_{O_2} is 40 mm Hg, which is the case in the tissue capillaries. Point A on the dashed curve shows that the normal P_{CO_2} of 45 mm Hg in the tissues causes 52 volumes per cent of carbon dioxide to combine with the blood. On entering the lungs the P_{CO_2} falls to 40 mm Hg while the P_{O_2} rises to 100 mm Hg. If the carbon dioxide dissociation curve did not shift because of the Haldane effect, the carbon dioxide content of the blood would fall only to 50 volumes per cent, which would be a loss of only 2 volumes per cent of carbon dioxide. However, the increase in P_{O_2} in the lungs decreases the carbon dioxide dissociation curve from the dashed curve to the solid curve of the figure, so that the carbon dioxide content falls to 48 volumes per cent (point B). This represents an additional 2 volumes per cent loss of carbon dioxide. Thus, the Haldane effect approximately doubles the amount of carbon dioxide released from the blood in the lungs and approximately doubles the pickup of PCO_2 in the tissues.

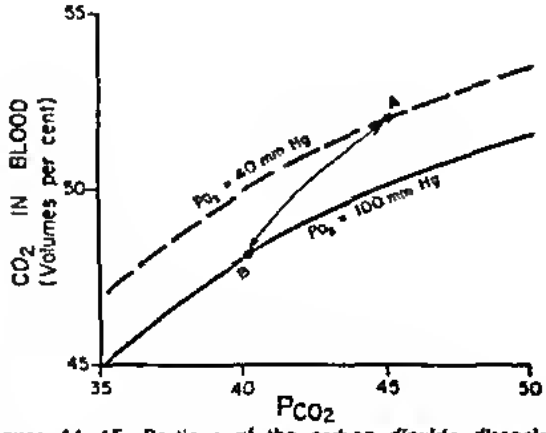


Figure 41–15. Portions of the carbon dioxide dissociation curves when the P_{O_2} is 100 mm Hg and 40 mm Hg, respectively. The arrow represents the Haldane effect on the transport of carbon dioxide, as explained in the text.

CHANGE IN BLOOD ACIDITY DURING CARBON DIOXIDE TRANSPORT

The carbonic acid formed when carbon dioxide enters the blood in the tissues decreases the blood pH. Fortunately, though, the reaction of this acid with the buffers of the blood prevents the hydrogen ion concentration from rising greatly. Ordinarily, arterial blood has a pH of approximately 7.41, and, as the blood acquires carbon dioxide in the tissue capillaries, the pH falls to approximately 7.37. In other words, a pH change of 0.04 unit takes place. The reverse occurs when carbon dioxide is released from the blood in the lungs, the pH rising to the arterial value once again. In exercise, or in other conditions of high metabolic activity, or when the blood flow through the tissues is sluggish, the decrease in pH in the tissue blood (and in the tissues themselves) can be as much as 0.50, or on occasion even more than this, thus causing tissue acidosis.

THE RESPIRATORY EXCHANGE RATIO

The discerning student will have noted that normal transport of oxygen from the lungs to the tissues by each deciliter of blood is about 5 ml, while normal transport of carbon dioxide from the tissues to the lungs is about 4 ml. Thus, under normal resting conditions only about 80 per cent as much carbon dioxide is expired from the lungs as there is oxygen uptake by the lungs. The ratio of carbon dioxide output to oxygen uptake is called the *respiratory exchange ratio* (R). That is,

$$R = \frac{\text{Rate of carbon dioxide output}}{\text{Rate of oxygen uptake}}$$

The value for R changes under different metabolic conditions. When a person is utilizing entirely carbohydrates for body metabolism, R rises to 1.00. On the other hand, when the person is utilizing fats almost entirely for metabolic energy, the level falls to as low as 0.7. The reason for this difference is that when oxygen is metabolized with carbohydrates one molecule of carbon dioxide is formed for each molecule of oxygen consumed, while when oxygen reacts with fats a large share of the oxygen combines with hydrogen atoms from the fats to form water instead of carbon dioxide. In other words, the *respiratory quotient of the chemical reactions* in the tissues when fats are metabolized is about 0.70 instead of 1.00 which is the case when carbohydrates are being utilized. The tissue respiratory quotient will be discussed in Chapter 71.

For a person on a normal diet consuming average amounts of carbohydrates, fats, and proteins, the average value for R is considered to be 0.825.

REFERENCES

Adamson, J. W, and Finch, C A. Hemoglobin function, oxygen affinity, and erythropoietin. *Annu. Rev. Physiol*, *37* 351, 1975

Bartels, H, and Baumann, R. Respiratory function of hemoglobin *Int. Rev. Physiol*, *14* 107, 1977

Bauer, C., *et al* (eds.) Biophysics and Physiology of Carbon Dioxide New York, Springer-Verlag, 1980

Bruley, D. F. Mathematical considerations for oxygen transport to tissue *Adv Exp Med. Biol*, *37* 749, 1973

Caughey, W. S (ed.) Oxygen, Biochemical and Clinical Aspects New York, Academic Press, 1979

Chance, B. Regulation of intracellular oxygen. *Proc Int Union Physiol. Sci*, *6* 13, 1968.

Cherniack, N S, and Longobardo, G S Oxygen and carbon dioxide gas stores of the body. *Physiol Rev*, *50* 196, 1970

Crowell, J W, and Smith, E E Determinants of the optimal hematocrit *J. Appl Physiol*, *22* 501, 1967

Forster, R. E. CO_2, chemical, biochemical, and physiological aspects. *Physiologist*, *13* 398, 1970

Forster, R. E, Pulmonary ventilation and blood gas exchange *In* Sodeman, W. A., *Jr, and Sodeman*, W A *(eds) Pathologic Physiology* Mechanisms of Disease, 5th Ed. Philadelphia, W B Saunders Company, 1974, p 371

Grodins, F. S, and Yamashiro, S M Optimization of the mammalian respiratory gas transport system. *Annu Rev Biophys Bioeng*, *4* 115, 1973.

Haldane, J. S, and Priestley, J G Respiration. New Haven, Yale University Press, 1935

Hayaishi, O Molecular Oxygen in Biology New York, American Elsevier Publishing Co, 1974

Johns, F. P.: Intracellular metabolism of oxygen *Am Rev Resp Dis*, *110* 58, 1974

Jones, C. E, *et al*., Determination of mean tissue oxygen tensions by implanted perforated capsules. *J Appl Physiol*, *26* 630, 1969

Jones, N. L. Blood Gases and Acid-Base Physiology New York, B C. Decker, 1980

Kessler, M, *et al* (eds.) Oxygen Supply Baltimore, University Park Press, 1973.

Kilmartin, J V, and Rossi-Bernardi, L. Interaction of hemoglobin with hydrogen ions, carbon dioxide, and organic phosphates *Physiol Rev*, *53* 836, 1973

Konigsberg, W Protein structure and molecular dysfunction. Hemoglobin. *In* Bondy, P K, and Rosenberg, L E (eds) Metabolic Control and Disease, 8th Ed. Philadelphia, W B Saunders Co, 1980, p 27

Lehman, H., and Huntsman, R G Man's Hemoglobins, 2nd Ed. Philadelphia, J B Lippincott Co, 1974

Maren, T H Carbonic anhydrase Chemistry, physiology, and inhibition *Physiol Rev*, *47* 595, 1967

Michel, C C: The transport of oxygen and carbon dioxide by the blood. *In* MTP International Review of Science Physiology Vol 2 Baltimore, University Park Press, 1974, p 67

Morson, B C (ed.) Hypoxia and Ischaemia London, Journal of Clinical Pathology for the College, 1977

Oski, F A., and Gottlieb, A J The interrelationships between red blood cell metabolites, hemoglobin, and the oxygen-equilibrium curve *Prog Hematol*, *7* 33, 1971

Painter, A. S, and Gill Kumar, P (eds.) Respiratory Adaptations, Capillary Exchange, and Reflex Mechanisms Delhi, India, Vallabbhbhai Patel Chest Institute, University of Delhi, 1977

Perutz, M F Hemoglobin structure and respiratory transport *Sci Am*, *239*(6) 92, 1978

Randall, D J The Evolution of Air Breathing in Vertebrates, New York, Cambridge University Press, 1980

Rubin, E D, and Simon, L. M Oxygen transport and cellular respiration *In* Frohlich, E D (ed) Pathophysiology, 2nd Ed. Philadelphia, J B Lippincott Co, 1976, p 167

Root, W S Carbon monoxide *In* Fenn, W O, and Rahn, H (eds) Handbook of Physiology Sec. 2, Vol 2 Baltimore, Williams & Wilkins Co 1965, p 1087

Stainsby, W N, and Barclay, J K Oxygen uptake by striated muscle Muscle Biol, *1* 273, 1972

Wagner, P D The oxyhemoglobin dissociation curve and pulmonary gas exchange *Sem Hematol*, *11* 405, 1974

Wagner, P D Diffusion and chemical reaction in pulmonary gas exchange *Physiol Rev* *57* 257, 1977

West, J B Pulmonary gas exchange *Int Rev Physiol*, *14* 83, 1977

Whalen, W J Intracellular PO_2 in heart and skeletal muscle *Physiologist*, *14* 69, 1971

Wittenberg, J B Myoglobin facilitated oxygen diffusion Role of myoglobin in oxygen entry into muscle *Physiol Rev*, *50* 559, 1970

42

Regulation of Respiration

The nervous system adjusts the rate of alveolar ventilation almost exactly to the demands of the body so that the blood oxygen pressure (P_{O_2}) and carbon dioxide pressure (P_{CO_2}) are hardly altered even during strenuous exercise or other types of respiratory stress.

The present chapter describes the operation of this neurogenic system for regulation of respiration.

THE RESPIRATORY CENTER

The "respiratory center" is composed of several widely dispersed groups of neurons located *bilaterally* in the medulla oblongata and pons, as illustrated in Figure 42–1. It is divided into three major collections of neurons. (1) a *dorsal respiratory group*, located in the dorsal portion of the medulla, which mainly causes inspiration, (2) a *ventral respiratory group*, located in the ventrolateral part of the medulla, which can cause either expiration or inspiration depending upon which neurons in the group are stimulated, and (3) the *pneumotaxic center*, located dorsally in the superior portion of the pons, which helps control both the rate and pattern of breathing. The dorsal respiratory group of neurons plays the fundamental r in the control of respiration. Therefore, let discuss its function first.

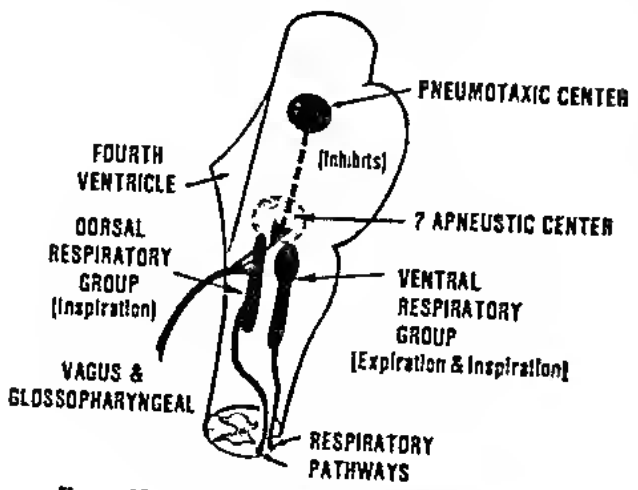


Figure 42–1. Organization of the respiratory center

The Dorsal Respiratory Group of Neurons

The dorsal respiratory group of neurons exter approximately the entire length of the medul Either all or most of its neurons are located with the *nucleus of the tractus solitarius*, though ad tional neurons in the adjacent reticular substan of the medulla probably also play important rol in respiratory control. The nucleus of tha tract solitarius is also the sensory termination of bo the vagal and glossopharyngeal nerves, whi transmit sensory signals into the respiratory ce ter from the peripheral chemoreceptors, the bar receptors, and several different types of recepto in the lung. All the signals from these peripher areas help in the control of respiration, as we sha discuss in subsequent sections of this chapter.

The Inspiratory Function of the Dorsal Res piratory Group of Neurons. Stimulation of th neurons in the dorsal respiratory group alway causes inspiration, never expiration. However, i will be recalled that normal quiet respiration i caused almost entirely, if not entirely, by contrac tion of the inspiratory muscles. Then expiration is caused passively by elastic recoil of the distended chest and lungs.

Rhythmical Inspiratory Discharges from the Dorsal Respiratory Group. The basic rhythm of respiration is generated in the dorsal respiratory group of neurons. Even when all the peripheral nerves entering the medulla are sectioned and also the brain stem is transected both above and below the medulla, this group of neurons still emits repetitive bursts of inspiratory action potentials. Unfortunately, though, the basic cause of these repetitive discharges is still unknown. In primitive animals, neural networks have been found in which activity of one set of neurons excites a second set, which in turn inhibits the first. Then after a period of time the mechanism repeats itself, continuing throughout the life of the animal.

Therefore, most respiratory physiologists believe that some similar network of neurons located within the medulla, probably involving not only the dorsal respiratory group but adjacent areas of the medulla as well, is responsible for the basic rhythm of respiration.

The Inspiratory "Ramp" Signal. The nervous signal that is transmitted to the inspiratory muscles is not an instantaneous burst of action potentials. Instead, in normal respiration, it begins very weakly at first and increases steadily in a ramp fashion for about 2 seconds. Then, it abruptly ceases for approximately the next 3 seconds, then begins again for still another cycle, and again and again. Thus, the inspiratory signal is said to be a *ramp signal*. The obvious advantage of this is that it causes a steady increase in the volume of the lungs during inspiration, rather than inspiratory gasps.

There are two ways in which the inspiratory ramp is controlled:

(1) Control of the rate of increase of the ramp signal, so that during very active respiration the ramp increases rapidly end therefore fills the lungs rapidly as well.

(2) Control of the limiting point at which the ramp suddenly ceases. This is the usual method for controlling the rate of respiration; that is, the earlier the ramp ceases the shorter the duration of inspiration, end for reasons not presently understood this also shortens the duration of expiration. Thus, the rate of respiration is increased.

The Pneumotaxic Center—Its Function in Limiting the Duration of Inspiration and Increasing Respiratory Rate

The pneumotaxic center, located dorsally in the *nucleus parabrachialis* of the upper pons, transmits impulses continuously to the inspiratory area. The primary effect of these is to control the "switch-off" point of the inspiratory ramp, thus controlling the duration of the filling phase of the lung cycle. When the pneumotaxic signals are strong, inspiration might last for as little as 0.5 second, but when weak, the inspiratory ramp might continue to rise for perhaps as long as 5 to 10 seconds, thus filling the lungs with a great excess of air.

Therefore, the function of the pneumotaxic center is primarily to limit inspiration. However, this has a secondary effect on the rate of breathing because limitation of inspiration also shortens the entire period of respiration. Thus, a strong pneumotaxic signal can increase the rate of breathing up to 30 to 40 breaths per minute, whereas a weak pneumotaxic signal may reduce the rate to only a few breaths per minute.

Limitation of Inspiration by Lung Inflation Signals—The Hering-Breuer Inflation Reflex

Located in the walls of the bronchi and bronchioles throughout the lungs are *stretch receptors* that transmit signals through the *vagi* into the dorsal respiratory group of neurons when the lungs become overstretched. These signals affect inspiration in much the same way as signals from the pneumotaxic center; that is, they limit the duration of inspiration. Therefore, when the lungs become overly inflated, the stretch receptors activate an appropriate feedback response that "switches-off" the inspiratory ramp and thus limits further inspiration. This is called the *Hering-Breuer inflation reflex*. This reflex also increases the rate of respiration because of the reduced period of inspiration, the same as is true for signals from the pneumotaxic center.

However, in human beings, the Hering-Breuer reflex probably is not activated until the tidal volume increases to greater than approximately 1.5 liters. Therefore, this reflex appears to be mainly a protective mechanism for preventing excess lung inflation rather than an important ingredient in the normal control of ventilation.

The Ventral Respiratory Group of Neurons—Its Function in Both Inspiration and Expiration

Located about 5 mm anterior and lateral to the dorsal respiratory group of neurons is still another group, the ventral respiratory group, that is important during more active breathing. These neurons, like the dorsal respiratory group, also are present in the entire length of the medulla and are found mainly in the *nucleus ambiguus* rostrally and the *nucleus retroambiguus* caudally. The function of this area differs from that of the dorsal respiratory group in several important ways:

(1) The neurons of the ventral respiratory group remain almost totally inactive during normal quiet respiration. Furthermore, there is no evidence that the ventral respiratory neurons participate in the basic rhythmic oscillation that controls respiration.

(2) When the respiratory drive for increased pulmonary ventilation becomes greater than normal, respiratory signals then spill over into the ventral respiratory neurons from the basic oscillating mechanism of the dorsal respiratory area. As a consequence, the ventral respiratory area then contributes its share to the respiratory drive as well.

(3) Electrical stimulation of some of the neurons in the ventral group causes inspiration whereas stimulation of others causes expiration. Therefore, these neurons contribute to both inspiration and expiration. However, they are especially important

in providing the powerful expiratory forces during expiration. Thus, this area operates more or less as an overdrive mechanism when high levels of pulmonary ventilation are required.

Possibility of an "Apneustic Center" in the Lower Pons

To add to the confusion of our knowledge about respiratory center function, there is another strange center in the lower part of the pons, called the *apneustic center*. However, its function can be demonstrated only when the vagus nerves to the medulla have been sectioned and when the connections from the pneumotaxic center have also been blocked by transecting the pons in its midregion. Then the apneustic center of the lower pons sends signals to the dorsal respiratory group of neurons that prevent the "switch-off" of the inspiratory ramp signal. Therefore, the ramp continues for as long as 10 to 20 seconds, thus greatly overfilling the lungs. Consequently, the lungs become inflated near maximum, and only occasional short expiratory gasps occur.

The function of the apneustic center might be to provide extra drive to inspiration, but the pneumotaxic center and the stretch signals from the vagi normally override this excessive inspiratory drive and therefore allow normal respiration.

CONTROL OF OVERALL RESPIRATORY CENTER ACTIVITY

Up to this point we have discussed the basic mechanisms for causing inspiration and expiration, but it is also important to know how the overall level of respiratory center activity is controlled to match the ventilatory needs of the body. For example, during very heavy exercise, the rates of oxygen utilization and carbon dioxide formation are often increased to as much as 20 times normal, requiring commensurate increases in pulmonary ventilation. This is achieved in two different ways:

(1) By feedback excitation of respiratory center activity in response to changes in chemical composition of the blood, especially its concentrations of carbon dioxide, hydrogen ions, and oxygen.

(2) By excitatory signals from other parts of the nervous system, such signals are especially important during exercise.

The major purpose of the remainder of this chapter is to discuss this control of ventilation in response to the needs of the body.

CHEMICAL CONTROL OF RESPIRATION

The ultimate goal of respiration is to maintain proper concentrations of oxygen, carbon dioxide, and hydrogen ions in the body fluid. It is therefore, that respiratory activity is highly sponsive to changes in any one of these.

Excess carbon dioxide or hydrogen ions af respiration mainly by excitatory effects on respiratory center itself, causing greatly increased strength of both the inspiratory and expiratory signals to the respiratory muscles. The increase in ventilation increases the of carbon dioxide from the blood, this also hydrogen ions from the blood because carbon dioxide also decreases the blood acid.

Oxygen, on the other hand, does not have significant *direct* effect on the respiratory of the brain in controlling respiration. acts either entirely or almost entirely on peripheral chemoreceptors located in the carotid and aortic bodies, and these in turn transmit appropriate nervous signals to the respiratory center for control of respiration.

Let us discuss first the stimulation of the respiratory center itself by carbon dioxide and hydrogen ions.

DIRECT CHEMICAL CONTROL OF RESPIRATORY CENTER ACTIVITY BY CARBON DIOXIDE AND HYDROGEN IONS

The Chemosensitive Area of the Respiratory Center. Thus far, we have discussed mainly three different areas of the respiratory center: the dorsal respiratory group of neurons, the ventral respiratory group, and the pneumotaxic center. However, it is believed that none of these are affected directly by changes in blood carbon dioxide concentration or hydrogen ion concentration. Instead, a very sensitive *chemosensitive* area, illustrated in Figure 42–2, is located bilaterally only a few mi-

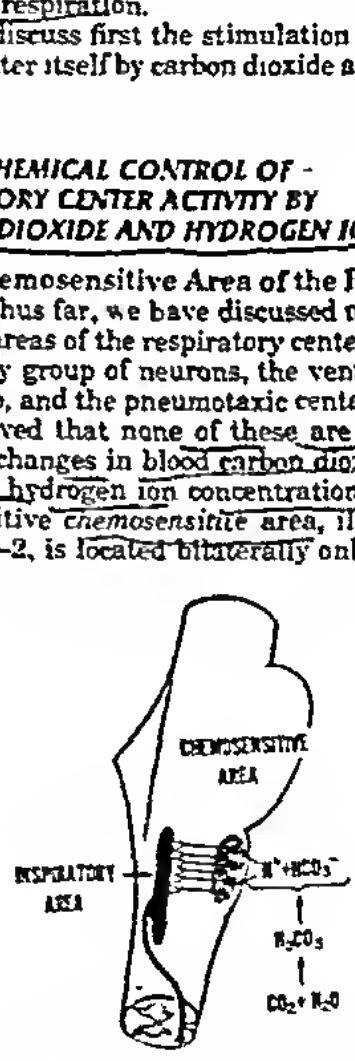


Figure 42–2. Stimulation of the inspiratory area by the chemosensitive area located bilaterally in the medulla, lying only a few microns beneath the ventral medullary surface. Note also that hydrogen ions stimulate the chemosensitive area, whereas mainly carbon dioxide in the fluid gives rise to the hydrogen ions.

rons beneath the surface of the medulla ventral to the entry of the glossopharyngeal and vagal nerves into the medulla. This area is highly sensitive to changes in either blood CO_2 or hydrogen ion concentration, and it in turn excites the other portions of the respiratory center. It has especially potent effects on increasing the degree of activity of the inspiratory center, increasing both the rate of rise of the inspiratory ramp signal and also the intensity of the signal. This in turn has an automatic secondary effect of increasing the frequency of the respiratory rhythm.

of the Chemosensitive Neurons to Hydrogen Ions—The Primary Stimulus

The sensor neurons in the chemosensitive area are especially excited by hydrogen ions; in fact, it is believed that hydrogen ions are perhaps the only important direct stimulus for these neurons. Unfortunately, though, hydrogen ions do not easily cross either the blood-brain barrier or the blood–cerebrospinal fluid barrier. For this reason, changes in hydrogen ion concentration in the blood actually have considerably less effect in stimulating the chemosensitive neurons than do changes in carbon dioxide, even though carbon dioxide stimulates these neurons indirectly, as will be explained below.

Effect of Blood Carbon Dioxide on Stimulating the Chemosensitive Area

Though carbon dioxide has very little direct effect to stimulate the neurons in the chemosensitive area, it does have a very potent indirect effect. It does this by reacting with the water of the tissues to form carbonic acid. This in turn dissociates into hydrogen and bicarbonate ions; the hydrogen ions then have a potent direct stimulatory effect. These effects are illustrated in Figure 42–2.

But, why is it that blood CO_2 has a more potent effect to stimulate the chemosensitive neurons than do blood hydrogen ions? The answer is that hydrogen ions, as noted above, pass through both the blood-brain barrier and the blood-cerebrospinal fluid barrier only very poorly whereas carbon dioxide passes through both these barriers almost as if they did not exist. Consequently, whenever the blood carbon dioxide concentration increases, so also does the P_{CO_2} in both the interstitial fluid of the medulla and also in the cerebrospinal fluid. And, in both of these fluids the carbon dioxide immediately reacts with the water to form hydrogen ions. Thus, paradoxically, more hydrogen ions are released into the respiratory chemosensitive sensory area when the blood carbon dioxide concentration increases than when the blood hydrogen ion concentration increases. For this reason, respiratory center activity is affected considerably more by changes in blood carbon dioxide than by changes in blood hydrogen ions, a fact that we will subsequently discuss quantitatively.

Stimulation of the Chemosensitive Area by Carbon Dioxide in the Cerebrospinal Fluid

Most of the stimulation of respiratory center activity probably results from changes in carbon dioxide and hydrogen ion concentrations in the interstitial fluid of the respiratory center itself. However, some respiratory physiologists believe that an increase in carbon dioxide concentration in the cerebrospinal fluid has a strong effect on stimulation of the chemosensitive area as well. This concept is still controversial but if it is true, it probably results from a difference in the degree of acid-base buffering of the interstitial versus the cerebrospinal fluid. The cerebrospinal fluid has very little protein buffer, whereas the interstitial fluid is lined on all sides by cells containing high concentrations of protein acid-base buffers. Therefore, a given change in carbon dioxide concentration in the cerebrospinal fluid will cause far more change in hydrogen ion concentration than in the interstitial fluid. And, since the chemosensitive neurons are located near the surface of the medulla, diffusion of hydrogen ions into these neurons from the cerebrospinal fluid could be a factor in the control of respiration.

One of the advantages of such a cerebrospinal fluid system for control of respiration would be the *rapidity* with which it could function. The cerebrospinal fluid is in intimate contact with the very rich blood supply of the arachnoid plexus. Therefore, within seconds after the blood P_{CO_2} changes, so also do the P_{CO_2} and hydrogen ion concentration of the cerebrospinal fluid change. On the other hand, several minutes are sometimes required for full change of the P_{CO_2} in the brain interstitial fluid.

Decrease in the Effect of Blood CO_2 on Respiratory Activity in One to Two Days. The stimulatory effect of increased carbon dioxide on respiration reaches its peak within a few minutes after an increase in blood P_{CO_2}. Thereafter, the stimulation gradually declines for the next one to two days to as little as one fifth the initial effect. The cause of this decreasing stimulation is not known, but one suggestion is that it results from readjustment of the bicarbonate ion concentration in the respiratory center or in the cerebrospinal fluid or both. The bicarbonate ions in turn would combine with the excess hydrogen ions, thus reducing the hydrogen ion concentration and simultaneously reducing the respiratory drive.

Therefore, a change in blood carbon dioxide concentration has a very potent *acute* effect on controlling respiration but only a weak *chronic* effect after a few days' adaptation.

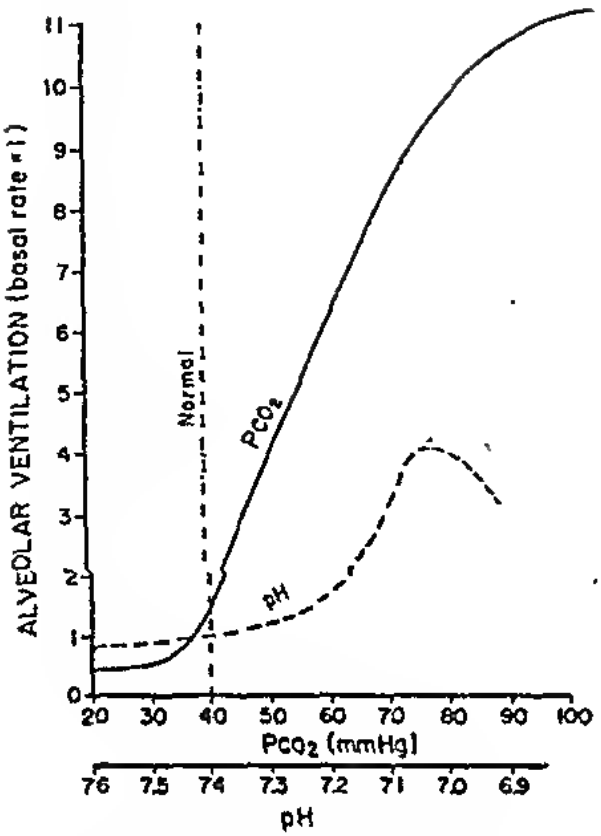


Figure 42–3. Effects of increased arterial P_{CO_2} and decreased arterial pH on the rate of alveolar ventilation.

Quantitative Effects of Blood P_{CO_2} and Hydrogen Ion Concentration on Alveolar Ventilation

Figure 42–3 illustrates quantitatively the approximate effects of blood P_{CO_2} and blood pH (which is an inverse measure of hydrogen ion concentration) on alveolar ventilation. Note the marked increase in ventilation caused by the increase in P_{CO_2}. But note also the much smaller effect of increased hydrogen ion concentration (that is, decreased pH).

Finally, note that this *difference* in stimulation of ventilation is especially great in the normal P_{CO_2} and pH ranges. P_{CO_2}s between 30 and 50 mm Hg and pHs between 7.5 and 7.3. Therefore, from a practical point, changes in blood carbon dioxide play by far the greater role in the normal minute-by-minute control of pulmonary ventilation.

Value of Carbon Dioxide as a Regulator of Alveolar Ventilation. Since carbon dioxide is one of the end-products of metabolism, its concentration in the body fluids greatly affects the chemical reactions of the cells and also affects the tissue pH. For these reasons, the tissue fluid P_{CO_2} must be regulated exactly, and, as pointed out in the preceding chapter, the respiratory system is the only effective means that the body has to control the blood and tissue P_{CO_2}. Therefore, stimulation of the respiratory center by carbon dioxide provides the necessary feedback mechanism for regulation of the concentration of carbon dioxide the body. That is, an increase in respiratory P_{CO_2} causes the respiratory system to el the excess carbon dioxide. Thus, the center maintains the P_{CO_2} of the tissue fluids relatively constant level and, therefore, might be called a "carbon dioxide pressostat."

THE PERIPHERAL CHEMORECEPTOR SYSTEM FOR CONTROL OF RESPIRATORY ACTIVITY—ROLE OF OXYGEN IN RESPIRATORY CONTROL

Aside from the direct sensitivity of the resp tory center itself to CO_2 and hydrogen ions, spe chemical receptors called *chemoreceptors*, loc outside the central nervous system, are also sponsive to changes in oxygen, carbon dioxide, hydrogen ion concentrations. These transmit nals to the respiratory center to help regu respiratory activity.

The chemoreceptors are located in the *car* and *aortic bodies*, which are illustrated in Fig 42–4 along with their afferent nerve connect to the respiratory center. The *carotid bodies* located bilaterally in the bifurcations of the c mon carotid arteries, and their afferent nerv bers pass through Hering's nerves to the *glo pharyngeal nerves* and thence to the do respiratory area of the medulla. The *aortic bo* are located along the arch of the aorta; t afferent nerve fibers pass through the *vagi* to dorsal respiratory area. Each of these chemore tor bodies receives a special blood supply thro a minute artery directly from the adjacent arte trunk.

Stimulation of the Chemoreceptors by creased Arterial Oxygen. Changes in arte oxygen concentration have *no* direct stimula

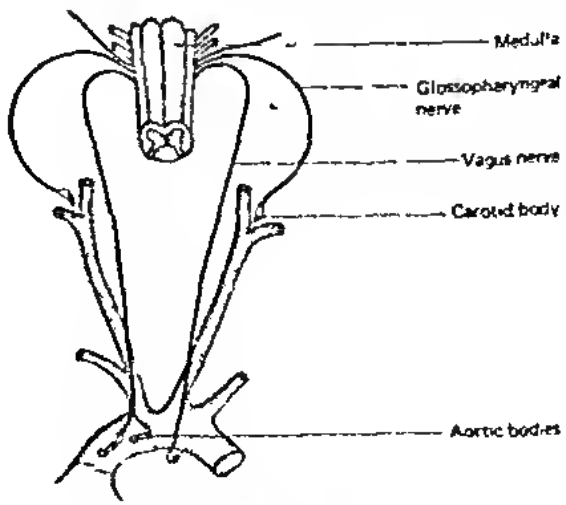


Figure 42–4. Respiratory control by the carotid and a bodies.

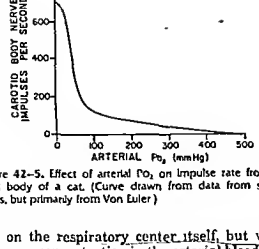


Figure 42–5. Effect of arterial Po_2 on impulse rate from the carotid body of a cat. (Curve drawn from data from several sources, but primarily from Von Euler.)

effect on the respiratory center itself, but when the oxygen concentration in the arterial blood falls below normal, the chemoreceptors become strongly stimulated. This effect is illustrated in Figure 42–5, which shows the relationship between *arterial* Po_2 and rate of nerve impulse transmission from a carotid body. Note that the impulse rate is particularly sensitive to changes in arterial Po_2 in the range between 60 and 30 mm Hg, which is the range in which the arterial hemoglobin saturation with oxygen decreases rapidly.

Effect of Carbon Dioxide and Hydrogen Ion Concentration on Chemoreceptor Activity. An increase in either carbon dioxide concentration or hydrogen ion concentration also excites the chemoreceptors and in this way indirectly increases respiratory activity. However, the direct effects of both these factors in the respiratory center itself are so much more powerful than their effects mediated through the chemoreceptors that for most practical purposes the indirect effects through the chemoreceptors do not need to be considered. In the case of oxygen, on the other hand, this is not true because diminished oxygen in the arterial blood can affect the respiration significantly *only* by acting through the chemoreceptors

Basic Mechanism of Stimulation of the Chemoreceptors by Oxygen Deficiency. The blood flow through the carotid and aortic bodies is extremely high, as high as that for almost any tissue in the body. Because of this, the A-V oxygen difference is less than 1 volume per cent, which means that the venous blood leaving the carotid bodies still has a Po_2 nearly equal to that of the arterial blood. It also means that the Po_2 of the tissues in the carotid and aortic bodies remains at all times almost equal to that of the arterial blood. Therefore, it is the *arterial* Po_2, not the venous Po_2, that *normally* determines the degree of stimulation of the chemoreceptors. Occasionally, though, serious hypotension can sufficiently decrease the blood flow through the carotid and aortic bodies to stimulate the chemoreceptors even when the arterial Po_2 is normal.

The exact means by which low Po_2 excites the nerve endings in the carotid and aortic bodies is still unknown. However, these bodies have two different, highly characteristic glandular-like cells in them. Therefore, some investigators have suggested that these cells might function as chemoreceptors and then in turn stimulate the nerve endings. However, other studies suggest that the nerve endings themselves are directly sensitive to the low Po_2.

Quantitative Effect of Low Blood Po_2 on Alveolar Ventilation

Low blood Po_2 normally will not increase alveolar ventilation significantly until the alveolar Po_2 falls almost to one half normal. This is illustrated in Figure 42–6. The lowermost curve of this figure shows that changing the alveolar arterial Po_2 from the normal value of slightly more than 100 mm Hg down to about 60 mm Hg has an imperceptible effect on ventilation. But, then, as the Po_2 falls still further, down to 40 and then to 30 mm Hg, alveolar ventilation increases 1.5– to 1.7-fold. However, contrast this rather feeble increase in alveolar ventilation to the fourfold increase caused by decreasing blood pH or the elevenfold increase caused by increasing the Pco_2. Thus, it is clear that the normal effect of changes in blood Po_2 on respiratory activity is very slight, especially when compared with the effect of Pco_2.

Cause of the Poor Response of Respiration to Low Po_2—The Opposing Effects of the Pco_2 and pH Regulatory Mechanisms. The cause of the poor effect of Po_2 changes on respiratory control is *opposition* caused by *both* the carbon dioxide and the hydrogen ion control mechanisms. This phenomenon can be explained by referring again to Figure 42–6. The increase in ventilation that does occur when the Po_2 falls blows off carbon dioxide from the blood and therefore decreases the Pco_2, which is also illustrated in the figure; at the same time it also decreases the hydrogen ion concentration. Therefore, two powerful respiratory inhibitory effects are caused: (a) diminished carbon dioxide and (b) diminished hydrogen ions. These two exert inhibitory effects that oppose the excitatory effect of the diminished oxygen. As a result, they keep the decreased oxygen from causing a marked increase in ventilation until the Po_2 falls to 20 to 40 mm Hg, a range that is incompatible with life for more than a few minutes. Therefore, the maximum effect of decreased alveolar oxygen on alveolar ventilation, in the range compatible with life, is normally only about a 66 per cent increase.

Thus, one can see that during normal respiration the Pco_2 and pH feedback control mechanisms are extremely powerful in relation to the Po_2 feedback control. Indeed, under normal conditions the Po_2

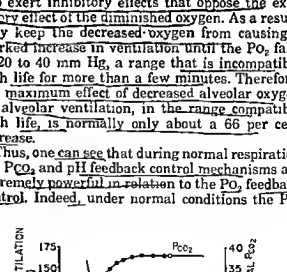


Figure 42–6. Effect of arterial Po_2 on alveolar ventilation and on the subsequent decrease in arterial Pco_2. (From Gray: Pulmonary Ventilation and Its Physiological Regulation. Charles C Thomas.)

mechanism is of almost no significance in the control of respiration.

Yet, under some abnormal conditions the P_{CO_2} and hydrogen ion concentrations *increase at the same time that the arterial P_{O_2} decreases.* Under these conditions, all three of the feedback mechanisms support each other, and the P_{O_2} mechanism then exerts its full share of respiratory stimulation, sometimes becoming even more potent as a controller of respiration than the P_{CO_2} and hydrogen ion mechanisms.

Effect of Diminished Oxygen on Alveolar Ventilation When the Carbon Dioxide and pH Mechanisms Do Not Oppose the P_{O_2} Effect. When the concentrations of carbon dioxide and hydrogen ions are prevented from changing while decreased blood P_{O_2} increases the ventilation, the effect of diminished P_{O_2} is then 8 to 10 times as *great as when the carbon dioxide and hydrogen* ion concentrations do change and oppose the P_{O_2} effect. This is illustrated in Figure 42-7 which compares the effects of acute changes in P_{O_2}, pH, and P_{CO_2} on alveolar ventilation when only one of these factors changes at a time while the other two are kept exactly constant. This figure illustrates that under these conditions changes in P_{O_2} have almost as much effect on alveolar ventilation as changes in the other two factors. The reason that the other two factors dominate normal respiratory control is that the P_{O_2} control mechanism does not exert a significant opposition effect on the P_{CO_2} and hydrogen ion mechanisms, in contrast to the powerful effect of these other two mechanisms in interfering with the P_{O_2} effect.

Conditions Under Which Diminished Oxygen Does Play a Major Role in the Regulation of Respiration. In pneumonia, emphysema, and other lung ailments in which gases are not readily exchanged between the atmosphere and the pulmonary blood, the oxygen regulatory system *does* then play a major role in the regulation of respiration. Contrary to the normal effect, the increased ventilation then caused by oxygen lack is not followed by reduced arterial P_{CO_2} and hydrogen ion concentration, because the pulmonary disease also diminishes carbon dioxide exchange as well as oxygen exchange. Instead, the CO_2 either remains constant or builds up in the blood, and the hydrogen ion concentration behaves similarly. Therefore, the opposition effects of these other two control systems on the oxygen lack system are absent. As a result, the oxygen lack stimulation develops its full power and can increase alveolar ventilation as much as five- to sevenfold.

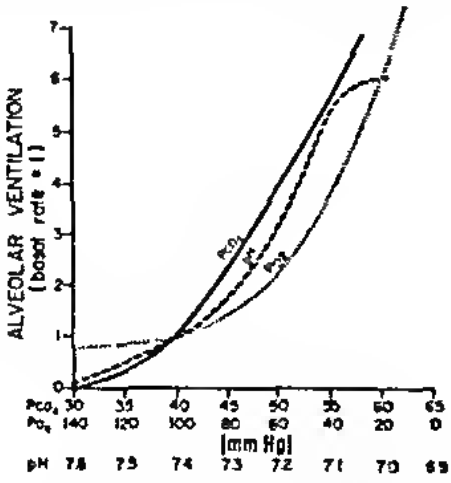


Figure 42-7. Approximate effects on alveolar ventilation of changing the concentrations of carbon dioxide, hydrogen ions, and oxygen in the arterial blood when only one of the humoral factors is changed at a time and the other two are maintained at absolutely normal levels.

Effects of the Oxygen Lack Mechanism at High Altitudes. When a person first ascends to high altitudes (or in any other way is exposed to a rarefied atmosphere), the diminished oxygen in *the air stimulates the oxygen lack control system* of respiration. The respiration at first increases to a maximum of about two thirds above normal, which is a comparatively slight increase. Once again, the cause of this slight increase is the tremendous opposition effects of the carbon dioxide and hydrogen ion control mechanisms on the oxygen lack mechanism.

However, over several days, the respiratory center gradually becomes "adapted" to the diminished carbon dioxide, as explained earlier in the chapter, so that its opposition effect to the oxygen control is gradually lost, and alveolar ventilation then rises to as high as five to seven times normal. This is part of the acclimatization that occurs as a person slowly ascends a mountain, thus allowing the person to adjust respiration gradually to a level fitted for the higher altitude.

Why Oxygen Regulation of Respiration Is Not Normally Needed. On first thought, it seems strange that oxygen should play so small a role in the normal regulation of respiration, particularly since one of the primary functions of the respiratory center is to provide adequate intake of oxygen. However, oxygen control of respiration is not needed under most normal circumstances for the following reason:

The respiratory system ordinarily maintains an alveolar P_{O_2} actually *higher* than the level needed to saturate almost completely the hemoglobin of the arterial blood. It does not matter whether alveolar ventilation is normal or ten times normal, the blood will still be essentially fully saturated. Also, alveolar ventilation can decrease to as low as one half normal, and the blood still remains within 10 per cent of complete saturation. Therefore, one can see that alveolar ventilation can change tremendously without significantly affecting oxygen transport to the tissues.

On the other hand, changes in alveolar ventilation do have a tremendous effect on both blood and tissue carbon dioxide concentration, as was explained earlier in the chapter. Therefore, it is exceedingly important that carbon dioxide—not

oxygen—be the major controller of respiration under normal conditions.

COMPOSITE EFFECTS OF P_{CO_2}, pH, AND P_{O_2} ON RESPIRATORY ACTIVITY

Now that the effects of each of the individual humoral factors on respiratory activity have been discussed, we need to see how all of them function together in controlling respiration. We have already seen that the different factors oppose each other in some instances but at other times support each other. One of the best attempts to show how these factors interact is illustrated in Figure 42–8, which shows two separate "families" of curves relating alveolar P_{CO_2} to pulmonary ventilation. The "family" of curves represented by the solid lines shows the effects of different levels of alveolar P_{CO_2} on alveolar ventilation determined for the average person at different alveolar P_{O_2}s between 40 and 100 mm Hg and with the pH constant at the normal level of 7.4. The dashed "family" of curves was determined after the same average person had developed metabolic acidosis, with a shift of the blood pH from 7.4 to 7.3. Both these families of curves show that the *slope* of the curve relating P_{CO_2} to alveolar ventilation increases markedly as the P_{O_2} decreases. To express this another way, a decrease in P_{O_2} in the alveoli *multiplies* the effects on alveolar ventilation of *changes* in carbon dioxide concentration.

Figure 42–8 shows also that acidosis shifts the entire family of curves to the left. This means that increased hydrogen ion concentration decreases the carbon dioxide concentration required to stimulate the respiratory center.

The curves in Figure 42–8 are important when one wishes to analyze respiratory control from a quantitative point of view because, in most instances in which respiratory control becomes altered, more than one of the individual humoral factors are changed at the same time.

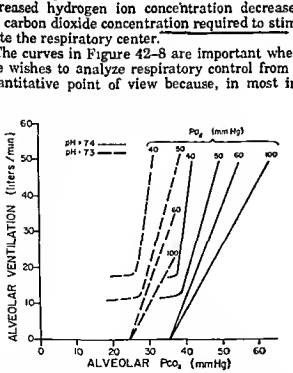


Figure 42–8. A composite diagram showing the interrelated effects of P_{CO_2}, P_{O_2}, and pH on alveolar ventilation. (Drawn from data presented in Cunningham and Lloyd The Regulation of Human Respiration. F. A. Davis Co.)

REGULATION OF RESPIRATION DURING EXERCISE

In strenuous exercise, oxygen utilization and carbon dioxide formation can increase as much as twentyfold. Yet, alveolar ventilation ordinarily increases almost exactly in step with the increased level of metabolism, as illustrated in Figure 42–9. Therefore, the blood P_{O_2}, P_{CO_2}, and pH all remain *almost exactly normal.*

In trying to analyze the factors that cause increased ventilation during exercise, one is tempted immediately to ascribe this to the chemical alterations in the body fluids during exercise, including increase of carbon dioxide, increase of hydrogen ions, and decrease of oxygen. However, this is not valid, for measurements of arterial P_{CO_2}, pH, and P_{O_2} show that none of these usually changes significantly and certainly not enough to account for more than a small percentage of the increase in ventilation. Indeed, even if a very high P_{CO_2} should develop during exercise, this still would be sufficient to account for only two thirds of the increased ventilation of heavy muscular exercise, for, as shown in Figure 42–10, the minute respiratory volume in exercise is about 50 per cent greater than that which can be effected by maximal carbon dioxide stimulation.

Therefore, the question must be asked: What is it during exercise that causes the intense ventilation? This question has not been answered, but at least two different effects seem to be predominantly concerned:

1. The brain, on transmitting impulses to the contracting muscles, is believed to transmit collateral impulses into the brain stem to excite the *respiratory center.* This is analogous to the stimulatory effect of the higher centers of the brain on

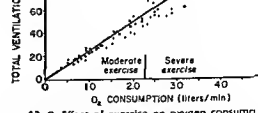


Figure 42–9. Effect of exercise on oxygen consump ventilatory rate (from Gray Pulmonary Ventilation and biological Regulation. Charles C Thomas.)

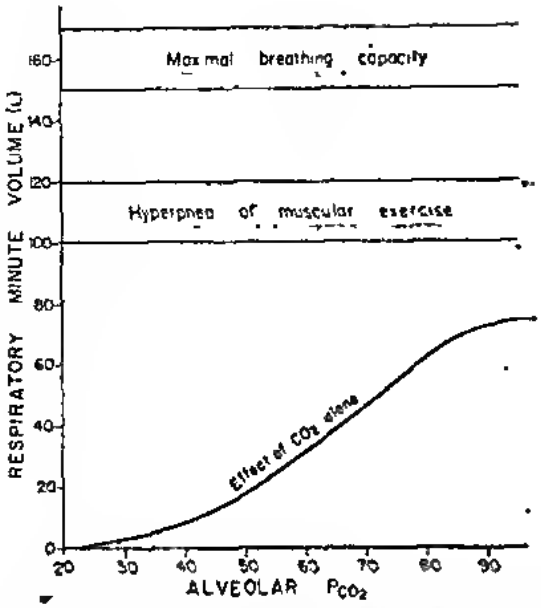


Figure 42–10. Relationship of hyperpnea caused by muscular exercise to that caused by increased alveolar P_{CO_2}. (Modified from Comroe: The Lung: Clinical Physiology and Pulmonary Function Tests. Chicago, Year Book Medical Publishers.)

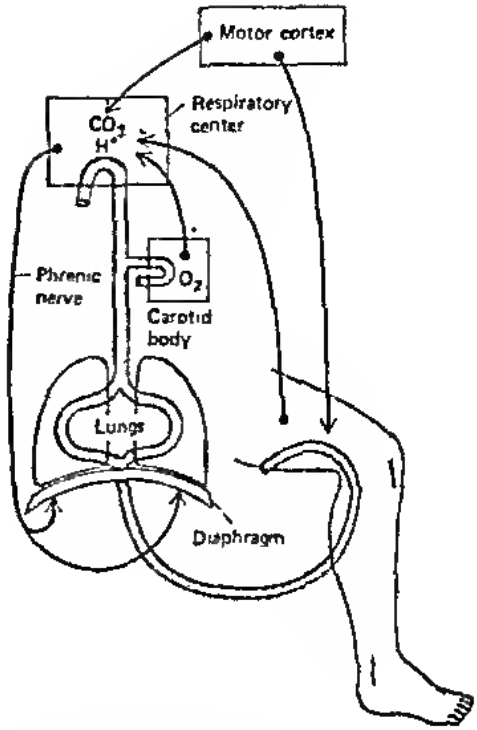


Figure 42–11. The different factors that enter into regulat of respiration during exercise.

the vasomotor center of the brain stem during exercise, causing a rise in arterial pressure as well as an increase in ventilation.

2. During exercise, the body movements, especially of the limbs, are believed to increase pulmonary ventilation by exciting joint proprioceptors that then transmit excitatory impulses to the respiratory center. The reason for believing this is that even passive movements of the limbs often increase pulmonary ventilation severalfold.

It is possible that still other factors are also important in increasing pulmonary ventilation during exercise. For instance, some experiments even suggest that hypoxia developing in the muscles during exercise elicits afferent nerve signals to the respiratory center to excite respiration. However, since the increase in ventilation begins immediately upon the initiation of exercise, most of the increase in respiration probably results from the two neurogenic factors noted above, namely *stimulatory impulses from the higher centers of the brain and proprioceptive stimulatory reflexes*.

Interrelationship Between Chemical Factors and Nervous Factors in the Control of Respiration During Exercise. Figure 42–11 illustrates the different factors that operate in the control of respiration during exercise, showing two neurogenic factors: (1) direct stimulation of the respiratory center by signals from the motor cortex, and (2) indirect stimulation by proprioceptors. It shows also the three humoral factors of carbon dioxide, hydrogen ions, and oxygen.

When a person exercises, usually the nervous factors stimulate the respiratory center almost exactly the proper amount to supply the extra oxygen requirements for the exercise and to blow off the extra carbon dioxide. But, occasionally, t nervous signals are either too strong or too we in their stimulation of the respiratory cent Then, the chemical factors play a very significa role in bringing about the final adjustment respiration required to keep the carbon dioxi and hydrogen ion concentrations of the body flui as nearly normal as possible. This effect is illu trated in Figure 42–12, which shows changes alveolar ventilation and in arterial P_{CO_2} during one-minute period of exercise and then for anoth

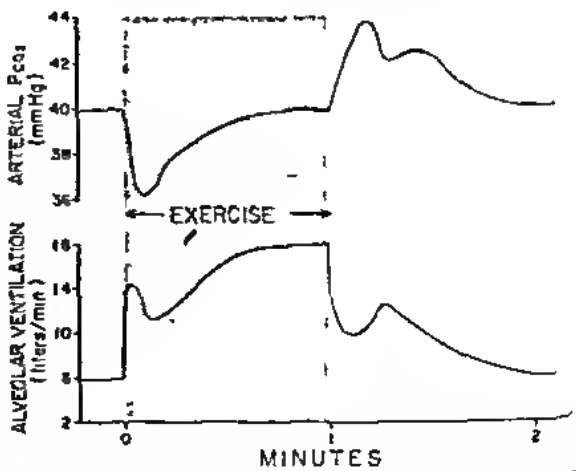


Figure 42–12. Changes in alveolar ventilation and arterial P_{C} during a 1-minute period of exercise and also following ter nation of the exercise. (Extrapolated to the human being fr data in dogs from Bainton. J. Appl. Physiol., 33:778, 1972.)

minute after the exercise is over. Note that the alveolar ventilation increases without an initial increase in arterial P_{CO_2}; this increase is caused by a stimulus that originates in the brain at the same time that the brain excites the muscles.

Often, the increase in alveolar ventilation at the onset of ventilation is so great that it actually *decreases* arterial P_{CO_2} below normal, as shown in the figure, even though the exercising muscles are beginning to form large amounts of carbon dioxide. The reason for this is that the ventilation forges ahead of the buildup of carbon dioxide in the blood. Thus, the brain provides an "anticipatory" stimulation of respiration at the onset of exercise, causing excessive alveolar ventilation even before this is needed. However, after about 30 to 40 seconds, the amount of carbon dioxide released into the blood from the muscles approximately matches the increased rate of ventilation. Then the arterial P_{CO_2} returns essentially to normal and no longer is a significant factor in controlling the ventilation. Thereafter, it is then mainly the nervous signals that drive the respiration, not the chemical signals.

To summarize, at the onset of exercise, an anticipatory signal from the brain initiates an immediate increase in alveolar ventilation; sometimes the increase is too much, sometimes too little. However, within less than one minute, the chemical feedback factors (mainly arterial P_{CO_2}) make additional adjustments, either upward or downward, to balance the rate of alveolar ventilation with the rate of metabolism in the body. Yet, on the average, because of the direct brain stimulus to respiration, the final levels of P_{CO_2}, H^+ concentration, and P_{O_2} in the arterial blood are very near normal during exercise—sometimes slightly above normal and sometimes slightly below normal.

Now for a final comment regarding the brain factor for stimulation of respiration during exercise: Many experiments indicate that this brain factor is partly if not entirely a learned response. That is, with repeated periods of exercise of the same degree of strenuousness, the brain seems to become progressively more able to establish the proper amount of brain signal to maintain the chemical factors at their normal levels during exercise. Therefore, there is much reason to believe that some of the higher centers of learning in the brain are important in this brain factor—probably even the cerebral cortex. Indeed, when a person is anesthetized, so that the higher brain centers become nonfunctional, respiration is then controlled almost entirely by the chemical factors.

OTHER FACTORS THAT AFFECT RESPIRATION

Voluntary Control of Respiration. Thus far, we have discussed the involuntary system for control of respiration. However, we all know that respiration can be controlled voluntarily, and that one can hyperventilate or hypoventilate to such an extent that serious derangements in P_{CO_2}, pH, and P_{O_2} can occur in the blood.

Voluntary control of respiration seems not to be mediated through control of the involuntary respiratory center of the medulla. Instead, the nervous pathway for voluntary control passes directly from the cortex and other higher centers downward through the corticospinal tract to the spinal neurons that drive the respiratory muscles.

Influence of Vasomotor Center Activity on Respiration. The vasomotor center that controls peripheral vasoconstriction and heart activity is closely related to the respiratory center in the medulla. Furthermore, a moderate degree of spillover of nerve signals occurs mutually between the two centers. Therefore, almost any factor that increases the activity of the vasomotor center also has at least a moderate effect on increasing respiration. For instance, a decrease in arterial pressure inhibits the baroreceptors, which increases vasomotor activity and increases pulmonary ventilation as well.

Effect of Body Temperature on Respiration. An increase in body temperature increases the rate of respiration. An *indirect* effect is caused by the action of the increased temperature in increasing cellular metabolism all through the body, which in turn enhances the chemical stimuli for increased respiration. Increased body temperature also has a *direct* effect by increasing respiratory center activity. Therefore, in patients with fever, both the rate and depth of ventilation are often greatly increased.

Panting. Many lower animals have very little ability to lose heat from the surfaces of their bodies for two reasons: First, *the surfaces are usually covered with fur*, and, second, the skin of most lower animals is not supplied with sweat glands, which prevents evaporative loss of heat. Therefore, a substitute mechanism, the *panting* mechanism, is used by many lower animals as a means for dissipating heat.

The phenomenon of panting is "turned on" by the hypothalamic thermoregulator center that will be discussed in Chapter 72. When the blood becomes overheated, the hypothalamus initiates neurogenic signals to decrease the body temperature. One of these signals is to initiate panting. The actual panting process is then controlled by a *panting center* that is closely related to the pneumotaxic center in the pons.

When an animal pants, it breathes in and out rapidly so that large quantities of new air from the exterior come in contact with the upper portions of its respiratory passages, and this cools the blood in the mucosa as a result of water evaporation from the mucosal surfaces. Yet, panting does not increase the alveolar ventilation more than is required for proper control of the blood gases, because each breath is extremely shallow; therefore, most of the air that enters the alveoli is dead space air.

Effect of Irritant Receptors in the Airways. The epithelium of the trachea, bronchi, and bronchioles is supplied with sensory nerve endings that are stimulated by irritants that enter the respiratory airways. These cause coughing and sneezing, as was discussed in Chapter 39. They possibly also cause bronchial constriction in such diseases as asthma and emphysema.

Function of Lung "J" Receptors. A few sensory nerve endings occur in the alveolar walls in *juxtaposition* to the pulmonary capillaries, from whence comes

the name "J" receptors. They are stimulated when irritant chemicals are injected into the pulmonary blood, and they are also excited when the pulmonary capillaries become engorged with blood or when pulmonary edema occurs in such conditions as congestive heart failure. Though the functional role of the J receptors is not known, their excitation perhaps does give the person a feeling of dyspnea.

ABNORMALITIES OF RESPIRATORY CONTROL

RESPIRATORY CENTER DEPRESSION

Cerebrovascular Disease. Probably the most common cause of long-term respiratory center depression is cerebrovascular disease in older-aged patients, especially following vascular occlusions or hemorrhages that damage the respiratory center areas. In such instances, a person may have chronically elevated arterial P_{CO_2}S and depressed P_{O_2}S.

Clinical Measurement of Respiratory Center Depression. In the clinical pulmonary function laboratory, the degree of respiratory center depression is frequently measured by the simple maneuver of having the patient breathe carbon dioxide at successively increasing concentrations. Then, from measurements of alveolar P_{CO_2} and alveolar ventilation, a *carbon dioxide-alveolar ventilation stimulation curve*, similar to that illustrated in Figure 42–3, is plotted. The degree of respiratory center depression is ascertained by comparing the slope of the carbon dioxide stimulation curve with that of the normal curve, assuming, of course, that all other aspects of pulmonary function are normal. If they are not normal, these also must be factored into the analysis.

Acute Brain Edema. The activity of the respiratory center may be depressed or even totally inactivated by acute brain edema resulting from brain concussion. For instance, the head might be struck against some solid object, following which the damaged brain tissues swell, compressing the cerebral arteries against the cranial vault and thus totally or partially blocking the cerebral blood supply. As a result, the neurons of the respiratory center first become inactive and later die. In this manner brain edema may either depress or totally inactivate the respiratory center.

Occasionally, respiratory depression resulting from brain edema can be relieved temporarily by intravenous injection of hypertonic solutions such as highly concentrated mannitol solution. These solutions osmotically remove some of the intracellular fluids of the brain, thus relieving intracranial pressure and sometimes re-establishing respiration within a few minutes.

Anesthesia. Perhaps the most prevalent cause of respiratory depression and respiratory arrest is *overdosage* of anesthetics or narcotics. The best agent to be used for anesthesia is one that depresses the respiratory center the least while depressing the cerebral cortex the [illegible]. Ether is among the best of the anesthetics by [illegible] criteria, though halothane, cyclopropane, ethyl[illegible], nitrous oxide, and a few others have almost the [illegible]me value. On the other hand, sodium pentobarbital is a poor anesthetic because it depresses the respiratory center considerably more than the above-mentioned agents. At one time, morphine was used as an anesthetic, but this drug is now used only as an adjunct to anesthetics because it greatly depresses the respiratory center.

PERIODIC BREATHING

An abnormality of respiration called *periodic breathing* occurs in a number of different disease conditions. The person breathes deeply for a short interval of time and then breathes slightly or not at all for an additional interval, the cycle repeating itself over and over again.

The most common type of periodic breathing, *Cheyne-Stokes breathing*, is characterized by slowly waxing and waning respiration, occurring over and over again every 45 seconds to 3 minutes.

Basic Mechanism of Cheyne-Stokes Breathing. Let us assume that the respiration becomes much more rapid and deeper than usual. This causes the P_{CO_2} in the pulmonary blood to *decrease*. A short time later the pulmonary blood reaches the brain, and the decreased P_{CO_2} begins to inhibit respiration. As a result, the pulmonary blood P_{CO_2} then *increases*. After another period of time the blood carrying the increased CO_2 arrives at the respiratory center and stimulates respiration again, thus making the person overbreathe once again and initiating a new cycle of depressed respiration. The cycles thus continue on and on, causing Cheyne-Stokes periodic breathing.

The successive changes in pulmonary and respiratory center P_{CO_2} during Cheyne-Stokes breathing are illustrated in Figure 42–13.

Since this basic feedback mechanism for causing Cheyne-Stokes respiration is present in every person, the real question that must be answered is not why Cheyne-Stokes breathing develops under certain conditions but, instead, why is it not present all the time in everyone? The reason for this is that the feedback is highly "damped," which may be explained as follows:

Damping of the Normal Respiratory Control Mechanism. In order for Cheyne-Stokes oscillation to occur, sufficient time must occur during the hyperpneic phase for the body fluid P_{CO_2} to fall considerably below the mean value. Then this decrease in P_{CO_2} initiates the apneic phase, which also must last long enough for the tissue P_{CO_2} to rise well above the mean value. In the normal person, the respiratory center ordinarily readjusts the breathing back toward normal before significant tissue P_{CO_2} changes can occur, thereby preventing the extreme excursions of P_{CO_2} in the tissue fluids that are necessary to cause a sustained oscillation. Thus, the tremendous storage capacity of the tissues for carbon

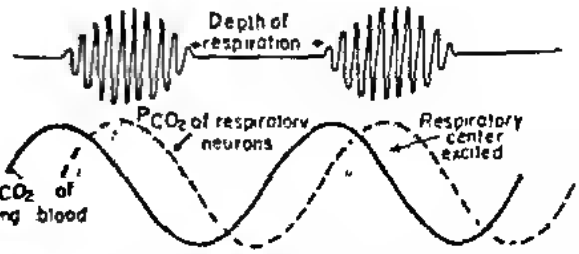


Figure 42–13. Cheyne-Stokes breathing, showing the changing P_{CO_2} in the pulmonary blood (solid line) and the delayed changes in P_{CO_2} of the fluids of the respiratory center (dashed line).

dioxide functions as an important "damping" factor to prevent Cheyne-Stokes respiration in the normal person.

A number of different conditions can overcome the damping of the feedback mechanism and cause it to oscillate spontaneously. Two of these are (1) increased delay time in the flow of blood from the lungs to the brain and (2) increased feedback gain of the respiratory control mechanisms.

Cheyne-Stokes Breathing Caused by Delay of Blood Flow from the Lungs to the Brain—Effect of Cardiac Failure. When the heart fails, blood flow from the lungs to the brain is sometimes greatly delayed, partly because the cardiac output decreases but mainly because the left heart sometimes becomes tremendous in size, thus requiring a long time for the blood to flow through the left ventricle and thence to the brain When such a delay occurs, the changes in P_{CO_2} caused by changed pulmonary ventilation are not detected by the brain for many seconds. Therefore, the blood P_{CO_2} continues to rise or fall to very abnormal levels before appropriate corrections are made in the ventilation.

Thus, a delay in transport of blood from the lungs to the brain is by itself sufficient cause for the development of Cheyne-Stokes breathing. In patients with chronic heart failure this type of breathing occasionally occurs continuously for months at a time.

Cheyne-Stokes Breathing Caused by Increased Feedback Gain in the Respiratory Center. Occasionally, damage to the brain stem increases the feedback gain of the respiratory center response to changes in blood P_{CO_2} and blood pH. The gain is defined as the *change* in ventilation caused by a given change in P_{CO_2} or pH. It is possible for the gain to increase while the sensitivity of the respiratory center decreases. For instance, let us assume that a *normal* P_{CO_2} of 40 mm Hg causes zero respiration, which represents greatly decreased sensitivity of the respiratory center to carbon dioxide. However, let us assume that in this same person an increase in P_{CO_2} to 45 mm Hg increases the ventilation to the normal rate. This represents a very high gain, because the per cent *increase* in ventilation is infinite. Consequently, the respiratory control system may then oscillate back and forth between apnea and breathing, which is the typical pattern of Cheyne-Stokes breathing. This effect probably explains why many patients with brain damage develop Cheyne-Stokes breathing even though at the same time they have depressed overall breathing; when this does occur it often is an ominous sign, foretelling that death is near.

Another instance in which increased gain of the respiratory control system causes Cheyne-Stokes breathing is during hypoxia. When a person becomes very hypoxic, the gain of the oxygen lack mechanism for controlling respiration increases as much as five- to tenfold, as was explained earlier in the chapter. When this extra gain is added to the gains of the carbon dioxide and pH mechanisms, typical Cheyne-Stokes oscillation often occurs. This explains the Cheyne-Stokes breathing that is often observed in mountain climbers, particularly at night while they are trying to sleep and the diminished respiration of sleep exacerbates the condition.

REFERENCES

Bainton, C R Effect of speed versus grade and shivering on ventilation in dogs during active exercise *J Appl. Physiol*, *33* 778, 1972

Biscoe, T. J : Carotid body Structure and function. *Physiol Rev*, *51* 427, 1971

Chernack, N S., and Longobardo, G S Cheyne-Stokes breathing An instability in physiologic control *N Engl J Med*, *288* 952, 1973

Cohen, M I Neurogenesis of respiratory rhythm in the mammal *Physiol Rev*, *59* 1105, 1979

Cohen, M I. Central determinants of respiratory rhythm *Annu Rev Physiol*, *43* 91, 1981

Coleridge, H M, *et al* Thoracic chemoreceptors in the dog A histological and electrophysiological study of the location, innervation and blood supply of the aortic bodies. *Circ Res*, *26* 235, 1970

Cunningham, D J C : Integrative aspects of the regulation of breathing A personal view *In* MTP International Review of Science Physiology Vol 2 Baltimore, University Park Press, 1974, p 303

Eldridge, F L, and Millhorn, D E Central regulation of respiration by endogenous neurotransmitters and neuromodulators *Annu. Rev Physiol*, *43* 121, 1981

Eyzaguirre, C, *et al* Arterial chemoreceptors. *In* Shepherd, J T, and Abbond, F M (eds.) Handbook of *Physiology Sec 2, Vol. III* Bethesda American Physiological Society, 1983 p 557

Fitzgerald, R., *et al* (eds.) The Regulation of Respiration During Sleep and Anesthesia New York. Plenum Press, 1978

Flenley, D C, and Warren, P M Ventilatory responses to O_2 and CO_2 during exercise *Annu Rev Physiol*, *45* 415, 1983

Guyton, A C *et al* Oscillating mechanism of Cheyne-Stokes breathing *Am J Physiol*, *187* 395, 1956

Haddad, G G, and Mellina, R B Hypoxia and respiratory control in early life *Annu Rev Physiol*, *46* 629, 1984

Hechtman, H B (ed.) Acute Respiratory Failure Etiology and Treatment West Palm Beach, Fla, CRC Press, 1979

Jansen, A H, and Chernick, V Development of respiratory control Physiol Rev, 63 437, 1983

Kalia, M P Anatomical organization of central respiratory neurons *Annu. Rev Physiol*, *43* 105, 1981

Karczewski, W A. Organization of the brainstem respiratory complex *In* MTP International Review of Science Physiology Vol 2 Baltimore, University Park Press, 1974, p 197

Loeschcke, H H Respiratory chemosensitivity in the medulla oblongata *Acta Neurobiol Exp*, *33* 297, 1973

Loeschcke, H H Central nervous chemoreceptors. *In* MTP International Review of Science Physiology Vol 2 Baltimore. University Park Press, 1974, p 167

Maruyama, H, and Honda, Y Differences in overall 'gain' of CO_2-feedback system between dead space and CO_2 ventilations in man. *Bull Eur Physiopathol Respir*, *20* 501, 1984.

Milhorn, H T, Jr, and Guyton, A. C. An analog computer analysis of Cheyne-Stokes breathing *J Appl Physiol*, *20* 328, 1965.

Milhorn, H T, Jr, *et al*. A mathematical model of the human respiratory control system *Biophys J*, *5* 27, 1965

Mitchell, G S, *et al* Changes in the VI-VCO_2 relationship during exercise in goats. Role of carotid bodies. *J Appl Physiol*, *57* 1894, 1984

Mitchell, R A, and Berger, A J Neural regulation of respiration *Am Rev Resp Dis*, *111* 206, 1975

Pack, A I Sensory inputs to the medulla. *Annu Rev Physiol*, *43* 73, 1981

Read, D J, and Henderson-Smart, D J Regulation of breathing in the newborn during different behavioral states. *Annu Rev Physiol*, *46* 675, 1984.

Rigatto, H. Control of ventilation in the newborn. *Annu Rev Physiol*, *46* 661, 1984

Schlachte, M E. (ed.). Central Neurone Environment and the Control Systems of Breathing and Circulation New York, Springer-Verlag, 1983

Von Euler, C, and Lagercrantz, H. (eds.) Central Nervous Control Mechanisms in Breathing New York, Pergamon Press, 1980

Walker, D W Peripheral and central chemoreceptors in the fetus and newborn *Annu Rev. Physiol*, *46* 687, 1984.

Whipp, B J : Ventilatory control during exercise in humans. *Annu Rev Physiol*, *45* 393, 1983

Williams, M. H (ed.) Symposium on Disturbances of Respiratory Control Philadelphia, W. B Saunders Co, 1980

43

Respiratory Insufficiency—Pathophysiology, Diagnosis, Oxygen Therapy

The diagnosis and treatment of most respiratory disorders have come to depend highly on an understanding of the basic physiological principles of respiration and gas exchange. Some diseases of respiration result from inadequate ventilation, while others result from abnormalities of diffusion through the pulmonary membrane or of transport from the lungs to the tissues. In each of these instances, the therapy is often entirely different so that it is no longer satisfactory simply to make a diagnosis of "respiratory insufficiency."

Definitions. In describing different respiratory disorders the following descriptive terms are used.

Eupnea means normal breathing, *tachypnea* means rapid breathing, and *bradypnea* means slow breathing.

Hyperpnea means a rate of alveolar ventilation great enough to cause over-respiration. However, this term is commonly used also to indicate simply a very high level of alveolar ventilation without necessarily implying over-respiration. *Hypopnea* is the opposite of hyperpnea, indicating under-respiration.

Anoxia means total lack of oxygen, but this term is more frequently used to mean simply decreased oxygen. A more correct term for decreased oxygen is *hypoxia*.

Anoxemia means lack of oxygen in the blood, but the term is generally used to mean simply reduced oxygen in the blood. A better term for this is *hypoxemia*.

Hypercapnia means excess carbon dioxide in the blood, and *hypocapnia* means depressed carbon dioxide. The term *acapnia* is also frequently used to imply hypocapnia, but, as is true with anoxemia, during life the absolute state of acapnia, with no carbon dioxide at all in the body fluids, never exists.

ADDITIONAL METHODS FOR STUDYING RESPIRATORY ABNORMALITIES

In the previous four chapters we have discussed a number of different methods for studying respiratory abnormalities, including measuring vital capacity, tidal air, functional residual capacity, dead space, physiologic shunt, physiologic dead space, maximum expiratory flow, forced expiratory volume, and respiratory center sensitivity. This array of measurements is only part of the armamentarium of the clinical pulmonary physiologist. Some other interesting tools available will be described.

Use of the Body Plethysmograph to Study Pulmonary Function. A body plethysmograph, which is illustrated in Figure 43–1, is simply an airtight tank into which a person is placed. It can be used in different ways to measure many different aspects of pulmonary function.

1. The body plethysmograph can be used as a spirometer. The person breathes through a tube to the exterior

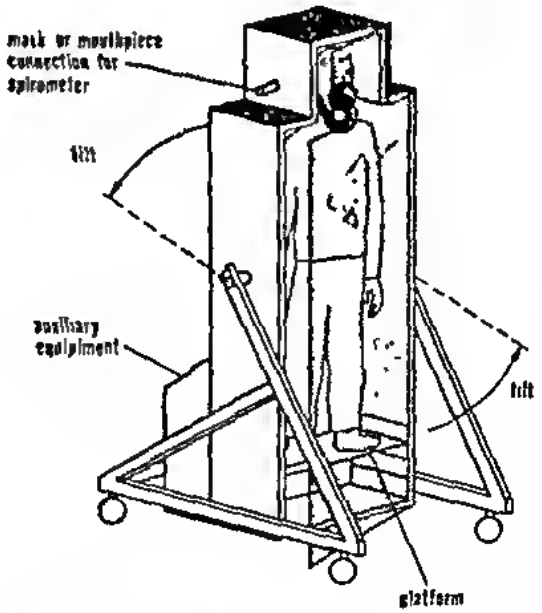


Figure 43–1. A body plethysmograph (from Ray [ed.], "Instrumentation for Pulmonary Function," in Medical Engineering © 1974 by Year Book Medical Publishers, Inc., Chicago. Used by permission. Adapted from "Pulmo Box," Med-Science Electronics, Inc., St. Louis, Missouri.)

of the tank. The excursions of the chest cause displacement of air volume within the tank, and an appropriate recording device connected to the tank can record this displacement in the form of volume changes, thus giving the usual spirometric recordings.

2. The body plethysmograph can be used to measure *lung volumes*. The person does not breathe to the exterior of the tank but instead breathes the air within the tank while the tank is completely closed, then expires with a forceful expiration against a manometer while the expiratory pressure is recorded. The increased pressure within the chest decreases the volume of gases in the chest in accordance with Boyle's law: *pressure times volume equals a constant*, thus the chest dimensions decrease. Therefore, the person's body requires less volume within the body plethysmograph. An appropriate calibrated pressure recorder connected to the tank records the amount of volume compression of the gases in the lungs. This measure, combined with the recording of pressure change within the chest, can be used to calculate the instantaneous volume of gas in the lung

3. The body plethysmograph can be used *to determine residual volume* of the lung The volume of air in the lung is measured by the above method and then as much volume as possible is expired. Subtracting the volume of the expired air from the previously measured lung volume gives the residual volume

4. The body plethysmograph can be used *to measure airway resistance*. The person is allowed to breathe the air within the tank while volume changes in the tank are recorded. During each expiration the gases in the lung become compressed because of the contraction of the chest wall so that the person's body during this brief period of time actually occupies less space within the chamber. This decreases the tank pressure, which is appropriately recorded. Furthermore, the amount of *decrease* in tank pressure is a measure of the pressure *increase* in the lungs, which in turn is proportional to the airway resistance times the expiratory air flow. (If the student will think through the logic of this measurement, as well as determine what other factors are needed to know to make the appropriate quantitative calculations of airway resistance, he or she will have an excellent test of the understanding of the mechanics of breathing.)

5 The body plethysmograph can be used *to measure pulmonary blood flow* The tank is completely closed from the surrounding atmosphere and the person breathes a gas mixture containing nitrous oxide from a flaccid balloon. As blood flows through the lung the nitrous oxide is absorbed into the pulmonary blood. This reduces the total quantity of gases in the tank (because the balloon itself is inside the tank) and thereby reduces its pressure. One can then calculate the rate of pulmonary blood flow from (a) the rate of reduction of the pressure in the tank, (b) the concentration of the nitrous oxide in the alveoli, and (c) the solubility coefficient of nitrous oxide in the blood.

Thus, the body plethysmograph is an extremely versatile instrument that has become an important tool for the clinical respiratory physiologist.

STUDY OF BLOOD GASES AND pH

Among the most fundamental of all tests of pulmonary performance are determinations of the blood Po_2, Pco_2, and pH. Indeed, it is often important to make these measurements extremely rapidly as an aid in determining appropriate therapy for acute respiratory distress or acute abnormalities of acid-base balance. Therefore, several simple and very rapid methods have been developed to make these measurements in only a few minutes, using no more than a few drops of blood. These methods are the following:

Determination of Blood pH. Blood pH is measured using a glass pH electrode of the type used in all chemical laboratories However, the electrodes used for this purpose are miniaturized so that no more than a drop or so of blood need be used. The voltage generated by the glass electrode is a direct measure of the pH, and this is generally read directly from a voltmeter scale, or it is recorded on a chart.

Determination of Blood Pco_2. A glass electrode pH meter can be used to determine blood Pco_2 in the following way When a weak solution of sodium bicarbonate is exposed to carbon dioxide, the carbon dioxide dissolves in the solution until an equilibrium state is established In this equilibrium state, the pH of the solution is a function of the carbon dioxide and bicarbonate ion concentrations in accordance with the Henderson-Hasselbalch equation that was discussed in Chapter 37 (that is, $pH = \log \frac{HCO_3}{CO_2} + 6.1$). Therefore, with appropriate calibration of the pH meter, the carbon dioxide concentration (or Pco_2) can be read directly from the meter

When this apparatus is used to measure Pco_2 in blood, a miniature glass electrode is coated with a thin solution of sodium bicarbonate, and this is separated from the blood by a very thin plastic membrane that allows carbon dioxide to diffuse from the blood into the solution. Here again, only a drop or so of blood is required

Determination of Blood Po_2. The concentration of oxygen in a fluid can be measured by a technique called *polarography*. Electrical current is made to flow between two electrodes in the solution If the voltage of the negative electrode is more than −0.6 volt different from the voltage of the solution, oxygen will deposit on the electrode Furthermore, the rate of current flow through the electrode will be directly proportional to the concentration of oxygen (and therefore Po_2 as well) In practice, a negative platinum electrode with a surface area of about 1 square millimeter is used, and this is separated from the blood by a thin plastic membrane that allows diffusion of oxygen but not diffusion of proteins or other substances that will "poison" the electrode. Once again, all parts of the system are miniaturized so that no more than a drop or so of blood is required.

Often, all three of the measuring devices for pH, Pco_2, and Po_2 are built into the same apparatus, and all these measurements can be made within a minute or so using a single, very small sample of blood. Thus, changes in the blood gases and pH can be followed almost moment by moment.

PHYSIOLOGIC TYPES OF RESPIRATORY INSUFFICIENCY

In general, the different types of abnormalities that cause respiratory insufficiency can be divided into three major categories: (1) those that cause *inadequate ventilation* of the alveoli, (2) those that reduce gaseous

diffusion through the respiratory membrane, and (3) those that decrease oxygen transport from the lungs to the tissues

ABNORMALITIES THAT CAUSE ALVEOLAR HYPOVENTILATION

The different respiratory abnormalities that can cause alveolar hypoventilation mainly have been discussed in the previous few chapters. However, let us summarize briefly some of the clinical causes of alveolar hypoventilation:

Severe hypoventilation frequently results from *paralysis of the respiratory muscles*, resulting from such factors as (1) *bulbar polio*, which often depresses the respiratory center, (2) *cervical transection of the spinal cord*, which prevents transmission of signals from the respiratory center to the respiratory muscles, or (3) *depression of the respiratory center* by anesthetics, drugs, or so forth.

Increased airway resistance is another very common cause of pulmonary hypoventilation; the two classic examples occur in *asthma* and *emphysema*.

Increased tissue resistance also often makes it difficult to ventilate the lungs. This can be caused by increased viscosity of the lung tissues resulting from such factors as (1) *emphysema*, (2) *pulmonary fibrosis*, (3) *tuberculosis*, (4) *various infections*, or (5) *pulmonary edema*.

Finally, another very common cause of alveolar hypoventilation is *decreased compliance of the lungs and chest wall*. As discussed in Chapter 39, any factor that makes it difficult for the lungs to expand will increase the "work" of ventilation and therefore also often will depress pulmonary ventilation. Some of the diseases of the lungs that do this include (1) *silicosis*, (2) *asbestosis*, (3) *sarcoidosis*, (4) *tuberculosis*, (5) *cancer*, and (6) *pneumonia*. In addition to the lung abnormalities that decrease compliance, restrictive diseases of the chest cage can also greatly decrease the compliance. These include especially *scoliosis* and *kyphosis*.

DISEASES THAT DECREASE LUNG DIFFUSING CAPACITY

Three different types of abnormalities can decrease the diffusing capacity of the lungs. These are: (1) decreased area of the respiratory membrane, (2) increased thickness of the respiratory membrane, called *alveolocapillary block*, and (3) abnormal ventilation-perfusion ratio in some parts of the lungs.

Decreased Area of the Respiratory Membrane. Diseases or abnormalities that decrease the area of the respiratory membrane include *removal of part or all of one lung*, *tuberculous destruction of the lung*, *cancerous destruction*, and *emphysema* which causes gradual destruction of alveolar septa.

Also, any acute condition that fills the alveoli with fluid or otherwise prevents air from coming in contact with the alveolar membrane, such as *pneumonia*, *pulmonary edema*, and *atelectasis*, can temporarily reduce the surface area of the respiratory membrane.

Increased Thickness of the Respiratory Membrane—Alveolocapillary Block. The most common acute cause of increased thickness of the respiratory membrane is *pulmonary edema* resulting from left heart failure or pneumonia. However, silicosis, *tuberculosis*, and *many other fibrotic conditions* can cause progressive deposition of fibrous tissue in the interstitial between the alveolar membrane and the [illegible] capillary membrane, thereby increasing the thickness of the respiratory membrane. This is usually called *alveolocapillary block*, or, occasionally, *interstitial fibrosis*. Since the rate of gaseous diffusion through the respiratory membrane is inversely proportional to the distance that the gas must diffuse, it is readily understood how alveolocapillary block can reduce the diffusing capacity of the lungs.

Abnormal Ventilation-Perfusion Ratio. In Chapter 40 we pointed out that the most common cause of decreased lung diffusing capacity is abnormal ventilation-perfusion ratio. That is, in some alveoli there is too little ventilation for the amount of blood flow so that the blood cannot become fully oxygenated. On the other hand, in other alveoli, ventilation is adequate but there is too little blood flow to accept the oxygen. Thus, in either instance, oxygen transfer to the blood becomes greatly compromised. These principles were discussed in detail in Chapter 40. It was pointed out that underventilated alveoli, in which $V_A/\dot{Q}$ is less than normal, leads to so-called *physiologic shunt*—that is, blood that is shunted past the lungs without becoming oxygenated. And it was also pointed out that the condition in which alveoli are overventilated while blood flow is too little—that is, $V_A/\dot{Q}$ is greater than normal—causes *physiologic dead space*. This means that there is ventilation that is not being used by the blood for oxygenation and therefore is *wasted ventilation*. It would be good for the student to return to Chapter 40 and review the principles of abnormal ventilation-perfusion ratio because of the very high prevalence of this abnormality in respiratory disease, especially in the most common of all serious lung diseases, *pulmonary emphysema* caused by smoking.

Diseases that cause abnormal ventilation-perfusion ratios include *thrombosis of a pulmonary artery*, *excessive airway resistance to some alveoli (emphysema)*, *reduced compliance of one lung without concomitant abnormality of the other lung*, and many other conditions that cause diffuse damage throughout the lungs.

ABNORMALITIES OF OXYGEN TRANSPORT FROM THE LUNGS TO THE TISSUES

Different conditions that reduce oxygen transport from the lungs to the tissues include *anemia*, in which the total amount of hemoglobin available to transport the oxygen is reduced, *carbon monoxide poisoning*, in which a large proportion of the hemoglobin becomes unable to transport oxygen, and *decreased blood flow to the tissues* caused by either low cardiac output or localized tissue ischemia. These have been discussed in Chapters 4, 41, and 28, respectively.

PHYSIOLOGIC PECULIARITIES OF SPECIFIC PULMONARY ABNORMALITIES

CHRONIC PULMONARY EMPHYSEMA

The term *pulmonary emphysema* literally means excess air in the lungs. However, when one speaks of chronic pulmonary emphysema, a complex destructive process of the lungs, caused in all but a few of the

patients by tobacco smoking, generally is meant. It results from three major pathophysiological events in the lung:

(1) *Chronic infection* caused by inhaling smoke or substances that irritate the bronchi and bronchioles. The principal reason for the chronic infection is that the irritant seriously deranges the normal protective mechanisms of the airways, including (a) partial paralysis of the cilia of the respiratory epithelium so that mucus cannot be moved easily out of the passageways, (b) stimulation of excess mucus secretion, which further exacerbates the condition, and (c) inhibition of the alveolar macrophages so that they become less effective in combating infection.

(2) The infection, the excess mucus, and inflammatory edema of the bronchiolar epithelium together cause *chronic obstruction* of many of the smaller airways.

(3) The obstruction of the airways makes it especially difficult to expire, thus causing *entrapment of air in the alveoli* and overstretching them. This, combined with the lung infection, causes marked destruction of many of the alveolar walls. Therefore, the final picture of the emphysematous lung is that illustrated in Figures 43–2 and 43–3.

The physiological effects of chronic emphysema are extremely varied, depending on the severity of the disease and on the relative degree of bronchiolar obstruction versus lung parenchymal destruction. However, among the different abnormalities are the following:

First, the bronchiolar obstruction greatly *increases airway resistance* and results in greatly increased work of breathing. It is especially difficult for the person to move air through the bronchioles during expiration because the compressive force on the outside of the lung not only compresses the alveoli but also compresses the bronchioles, which further increases their resistance during expiration.

Second, the marked loss of lung parenchyma greatly *decreases the diffusing capacity* of the lung, which reduces the ability of the lungs to oxygenate the blood and to remove carbon dioxide

Third, the obstructive process is frequently much worse in some parts of the lungs than in other parts so that some portions of the lungs are well ventilated while other portions are poorly ventilated. This often causes *extremely abnormal ventilation-perfusion ratios*, with a very low Va/Q in some parts *(physiological shunt)* re-

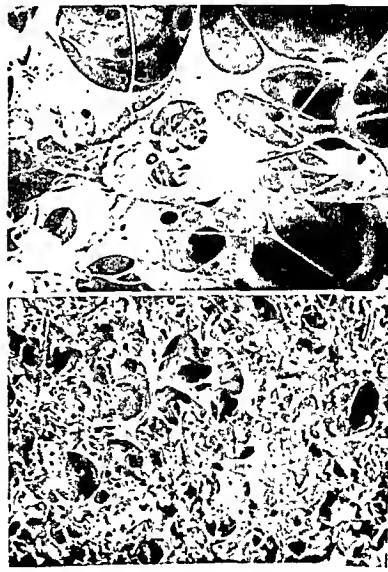


Figure 43–2. Contrast of the emphysematous lung (above) with the normal lung (below), showing extensive alveolar destruction. (Reproduced with permission of Patricia Delaney and the Department of Anatomy, The Medical College of Wisconsin.)

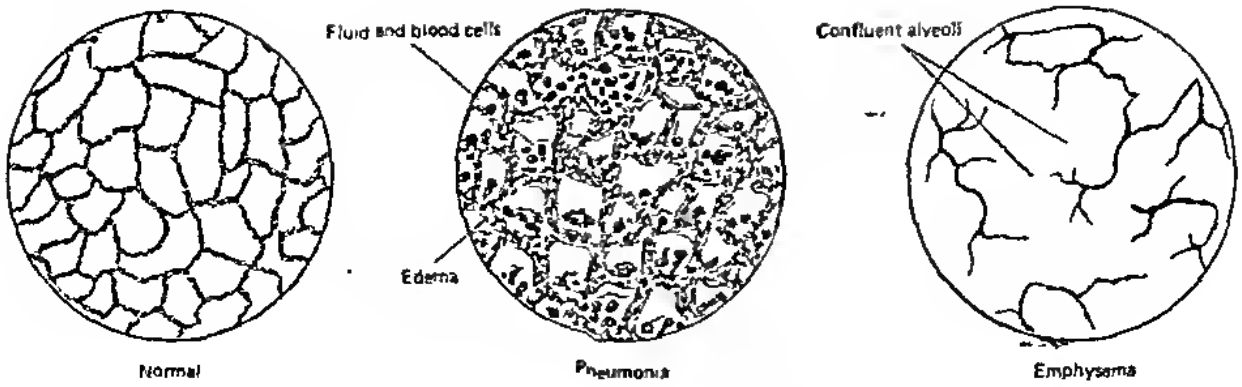


Figure 43–3. Pulmonary changes in pneumonia and emphysema.

sulting in poor aeration of the blood, and a very high Va/Q in other parts *(physiological dead space)* resulting in wasted ventilation, both effects occurring in the same lungs.

Fourth, loss of large portions of the lung parenchyma also decreases the number of pulmonary capillaries through which blood can pass. As a result, the pulmonary vascular resistance increases markedly, causing pulmonary hypertension. This in turn overloads the right heart and frequently causes right-heart failure.

Chronic emphysema usually progresses slowly over many years. The person develops hypoxia and hypercapnia because of hypoventilation of many alveoli and because of loss of lung parenchyma. The net result of all of these effects is severe and prolonged air hunger that can last for years until the hypoxia and hypercapnia cause death—a very high penalty to pay for smoking.

PNEUMONIA

The term pneumonia describes any inflammatory condition of the lung in which the alveoli are usually filled with fluid and blood cells, as shown in Figure 43–3. A common type of pneumonia is *bacterial pneumonia*, caused most frequently by pneumococci. This disease begins with infection in the alveoli, the pulmonary membrane becomes inflamed and highly porous so that fluid and even red and white blood cells pass out of the blood into the alveoli. Thus, the infected alveoli become progressively filled with fluid and cells, and the infection spreads by extension of bacteria from alveolus to alveolus. Eventually, large areas of the lungs, sometimes whole lobes or even a whole lung, become "consolidated," which means that they are filled with fluid and cellular debris.

The pulmonary function of the lungs during pneumonia changes in different stages of the disease. In the early stages, the pneumonia process might well be localized to only one lung, and alveolar ventilation may be reduced even though blood flow through the lung continues normally. This results in two major pulmonary abnormalities: (1) reduction in the total available surface area of the respiratory membrane and (2) decreased ventilation-perfusion ratio. Both these effects cause reduced diffusing capacity, which results in hypoxemia.

Figure 43–4 illustrates the effect of the decreased ventilation-perfusion ratio in pneumonia, showing that the blood passing through the aerated lung becomes 97 per cent saturated while that passing through the unaerated lung remains only 60 per cent saturated, causing the mean saturation of the aortic blood to be about 78 per cent, which is far below normal.

ATELECTASIS

Atelectasis means collapse of the alveoli. It can occur in a localized area of a lung, in an entire lobe, or in an entire lung. Its most common causes are two-fold: (1) obstruction of the airway, or (2) lack of surfactant in the fluids lining the alveoli.

Airway Obstruction. The airway obstruction type of atelectasis usually results from (a) blockage of many small bronchi with mucus or (b) obstruction of a major bronchus by either a large mucous plug or some solid object such as cancer. The air entrapped beyond the block is absorbed within minutes to hours by the blood flowing in the pulmonary capillaries. If the lung tissue is pliable enough, this will lead simply to collapse of the alveoli. However, if the lung tissue cannot collapse, absorption of air from the alveoli creates tremendously negative pressures within the alveoli and pulls fluid out of the pulmonary interstitium into the alveoli, thus causing the alveoli to fill completely with edema fluid. This almost always is the effect that occurs when an entire lung becomes atelectatic, a condition called *massive collapse* of the lung, because the solidity of the chest wall and of the mediastinum allows the lung to decrease only to about one half normal size rather than collapsing completely. The remainder of the space in the alveoli must then become filled with fluid.

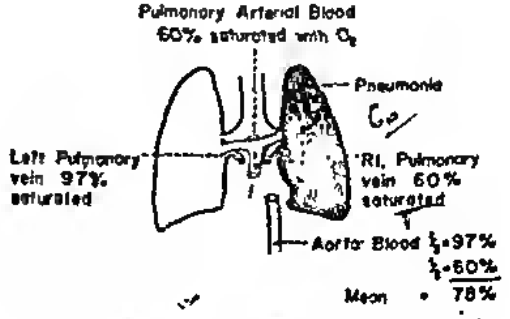


Figure 43–4. Effect of pneumonia on arterial blood oxygen saturation.

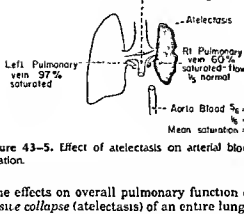


Figure 43–5. Effect of atelectasis on arterial blood oxygen saturation.

The effects on overall pulmonary function caused by *lassive collapse* (atelectasis) of an entire lung are illustated in Figure 43–5. Collapse of the lung tissue not ıly occludes the alveoli but also increases the resisance to blood flow through the pulmonary vessels This ›sistance increase occurs partially because of the colpse itself, which compresses and folds the vessels as ıe volume of the lung decreases But, in addition, ypoxia in the collapsed alveoli causes additional vasoinstriction, as was explained in Chapter 24

Because of the vascular constriction, blood flow ırough the atelectatic lung becomes slight. Most of the lood is routed through the ventilated lung and therefore ecomes well aerated. In the situation shown in Figure 3–5, five sixths of the blood passes through the aerated ıng and only one sixth through the unaerated lung As result, the overall ventilation-perfusion ratio is only ıoderately compromised, so that the aortic blood has nly mild oxygen desaturation.

Lack of Surfactant. In Chapter 39 it was pointed ut that the substance surfactant is secreted by the lveolar epithelium into the fluids that line the alveoli. his substance decreases the surface tension in the lveoli two- to tenfold. In the normal lung this plays a ıajor role in preventing alveolar collapse However, in number of different conditions the quantity of surfacant secreted by the alveoli is greatly decreased Sometimes this is severe enough to cause atelectasis For nstance, in the condition called *hyaline membrane disase* or *respiratory distress syndrome*, which often occurs n newborn babies, the quantity of surfactant secreted y the alveoli is greatly depressed This effect causes a erious tendency for the lungs of these babies to collapse, r to become filled with fluid as explained earlier, many f the infants die of suffocation as increasing portions of he lungs become atelectatic.

Atelectasis also frequently results from decreased surfactant secretion following extracorporeal perfusion of he lungs during open-heart surgery. The reason for the educed surfactant secretion in this instance is not nderstood.

ASTHMA

Asthma is characterized by spastic contraction of the bronchioles, which causes extremely difficult breathing t occurs in about 3 per cent of all persons at some time n life. The usual cause is hypersensitivity of the bronchioles to foreign substances in the air. In younger patients, under the age of 30 years, the asthma in about 70 per cent is caused by allergic hypersensitivity, especially sensitivity to plant pollens In older persons, the cause in about 70 per cent is hypersensitivity to nonallergic types of irritants in the air, such as irritants in smog.

The allergic reaction that occurs in the allergic type of asthma is believed to occur in the following way: The typically allergic person has a tendency to form abnormally large amounts of IgE antibodies, and these antibodies cause allergic reactions when they react with their complementary antigens, as was explained in Chapter 6. In asthma, these antibodies mainly attach to mast cells that lie in the lung interstitium in close association with the bronchioles and small bronchi. Then when the person breathes a pollen to which he or she is sensitive (that is, to which the person has developed IgE antibodies), the pollen reacts with the mast cell–attached antibodies and causes these cells to release several different substances Among them are *histamine, slow-reacting substance of anaphylaxis, eosinophilic chemotactic factor*, and *bradykinin* The combined effects of all these factors, especially of the slow-reacting substance of anaphylaxis (SRS-A), *are to produce (1) localized* edema in the walls of the small bronchioles as well as secretion of thick mucus into the bronchiolar lumens, and (2) spasm of the bronchiolar smooth muscle Obviously, therefore, the airway resistance increases greatly.

As discussed in Chapter 39, the bronchiolar diameter becomes more reduced during expiration than during inspiration in asthma The reason for this is that the increased intrapulmonary pressure during expiratory effort not only compresses the air in the alveoli but compresses the outsides of the bronchioles as well. Since the bronchioles are already partially occluded, further occlusion resulting from the external pressure creates especially severe obstruction during expiration. Therefore, the asthmatic person usually can inspire quite adequately but has great difficulty expiring Therefore, clinical measurements show greatly reduced maximum expiratory rate and timed expiratory volume. Also, this results in dyspnea, or "air hunger," which is discussed later in the chapter.

The functional residual capacity and the residual volume of the lung become greatly increased during the asthmatic attack because of the difficulty in expiring air from the lungs Over a long period of time the chest cage becomes permanently enlarged, causing a "barrel chest," and the functional residual capacity and residual volume also become permanently increased

TUBERCULOSIS

In tuberculosis the tubercle bacilli cause a peculiar tissue reaction in the lungs including, first, invasion of the infected region by macrophages and, second, walling off of the lesion by fibrous tissue to form the so-called "tubercle." This walling-off process helps to limit further transmission of the tubercle bacilli in the lungs and, therefore, is part of the protective process against the infection. However, in approximately 3 per cent of all persons who contract tuberculosis, the walling-off process fails, and tubercle bacilli spread throughout the lungs. Thus, tuberculosis in its late stages causes many areas of fibrosis throughout the lungs, and, secondly, it reduces the total amount of functional lung tissue. These effects cause (1) increased "work" on the part of the respiratory muscles to cause pulmonary ventilation and *reduced vital capacity and breathing capacity*, (2) *reduced*

total respiratory membrane surface area and *increased thickness of the respiratory membrane*, these causing progressively diminished pulmonary diffusing capacity, and (3) *abnormal ventilation-perfusion ratio* in the lungs, further reducing the pulmonary diffusing capacity.

HYPOXIA

Obviously, almost any of the conditions discussed in the past few sections of this chapter can cause serious degrees of cellular hypoxia. In some of these, oxygen therapy is of great value; in others it is of moderate value; in still others it is of almost no value. Therefore, it is important to classify the different types of hypoxia, then we can readily discuss the physiological principles of therapy. The following is a descriptive classification of the different causes of hypoxia.

1. Inadequate oxygenation of the lungs because of extrinsic reasons
   a. Deficiency of oxygen in atmosphere
   b. Hypoventilation (neuromuscular disorders)
2. Pulmonary disease
   a. Hypoventilation due to increased airway resistance or decreased pulmonary compliance
   b. Uneven alveolar ventilation-perfusion ratio (including increased physiologic dead space and physiologic shunt)
   c. Diminished respiratory membrane diffusion
3. Venous-to-arterial shunts ("right-to-left" cardiac shunts)
4. Inadequate transport and delivery of oxygen
   a. Anemia, abnormal hemoglobin
   b. General circulatory deficiency
   c. Localized circulatory deficiency (peripheral, cerebral, coronary vessels)
   d. Tissue edema
5. Inadequate tissue capability of using oxygen
   a. Poisoning of cellular enzymes
   b. Diminished cellular enzymes because of vitamin B deficiency

This classification of the different types of hypoxia is mainly self-evident from the discussions earlier in the chapter. Only one of the types of hypoxia in the above classification needs further elaboration; this is the hypoxia caused by inadequate capability of using oxygen. (Another classification of hypoxia that the student needs to know because it is widely used, even though far less precise than the preceding classification, is the following: (1) *Hypoxic hypoxia*, which includes all types of hypoxia in which not enough oxygen gets to the alveoli. (2) *Anemic hypoxia*, which is caused by inadequate hemoglobin or abnormal hemoglobin so that not enough oxygen can be transported to the tissues. (3) *Stagnant hypoxia* (or *circulatory hypoxia*), which means that there is not enough blood flow to carry the oxygen to the tissues. And (4) *histotoxic hypoxia*, which means that the tissues cannot utilize the oxygen even though it reaches the tissues in adequate quantities.)

Inadequate Tissue Capability of Using Oxygen. The classic cause of inability of the tissues to use oxygen is *cyanide poisoning*, in which the action of cytochrome oxidase is completely blocked—to such an extent that the tissues simply cannot utilize the oxygen even though plenty is available. Also, deficiencies of oxidative enzymes or other elements in the tissue oxidative system can lead to this type of hypoxia. A special example of this occurs in the disease beriberi, in which several important steps in the tissue utilization of oxygen and the formation of carbon dioxide are compromised because of vitamin B deficiency.

Effects of Hypoxia on the Body. Hypoxia, if severe enough, can actually cause death of the cells, but in less severe degrees it results principally in (1) depressed mental activity, sometimes culminating in coma, and (2) reduced work capacity of the muscles. These effects are discussed in the following chapter in relation to high altitude physiology and, therefore, are only mentioned here.

CYANOSIS

The term "cyanosis" means blueness of the skin, and its cause is excessive amounts of deoxygenated hemoglobin in the skin blood vessels, especially in the capillaries. This deoxygenated hemoglobin has an intense dark blue-purple color that is transmitted through the skin.

The presence of cyanosis is one of the most common clinical signs of different degrees of respiratory insufficiency, and for this reason it is important to understand the factors that determine the degree of cyanosis.

The most important factor determining the degree of cyanosis is the *quantity of deoxygenated hemoglobin in the arterial blood*. It is *not* the percentage deoxygenation of the hemoglobin that causes the bluish hue of the skin but principally the *concentration of deoxygenated hemoglobin without regard to the concentration of oxygenated hemoglobin*. The reason for this is that the red color of oxygenated blood is weak in comparison with the dark blue color of deoxygenated blood. Therefore, when the two are mixed together, the oxygenated blood has relatively little coloring effect in comparison with that of the deoxygenated blood.

In general, definite cyanosis appears whenever the *arterial blood* contains more than 5 grams per cent of deoxygenated hemoglobin. In polycythemia, cyanosis is very common because of the large amount of hemoglobin in the blood whereas in anemia, cyanosis is rare because it is difficult for there to be enough deoxygenated hemoglobin to produce the cyanotic color—even though anemia is much more likely to cause tissue hypoxia than is polycythemia.

Two other factors that affect the degree of cyanosis are (1) the rate of blood flow through the skin, and (2) the skin thickness. When the rate of flow is low, the capillary blood often loses most of its oxygen to the tissues and therefore becomes very blue even when the hemoglobin is fully oxygenated in the arterial blood. And the thinner the skin the easier is it to see the blue color of the capillary blood. For instance, in newborn babies, who have very thin skin, cyanosis occurs readily, particularly in highly vascular portions of the body such as the heels. Also, in adults as well as babies the lips and fingernails often appear cyanotic before the remainder of the body shows any blueness.

DYSPNEA

Dyspnea means a mental anguish associated with inability to ventilate enough to satisfy the demand for air. A common synonym is "air hunger."

At least three different factors often enter into the

levelopment of the sensation of dyspnea. These are (1) abnormality of the respiratory gases in the body fluids, especially hypercapnia (which will be discussed below) and to much less extent hypoxia, (2) the amount of work that must be performed by the respiratory muscles to provide adequate ventilation, and (3) the state of the mind.

A person becomes very dyspneic especially from excess buildup of carbon dioxide in body fluids. At times, however, the levels of both carbon dioxide and oxygen in the body fluids are completely normal, but to attain this normality of the respiratory gases, the person has to breathe forcefully. In these instances the forceful activity of the respiratory muscles frequently gives the person a sensation of very severe dyspnea.

Finally, the person's respiratory functions may be completely normal, and still dyspnea may be experienced because of an abnormal state of mind. This is called *neurogenic dyspnea*, or, sometimes, *emotional dyspnea*. For instance, almost anyone momentarily thinking about the act of breathing may suddenly start taking breaths a little more deeply than ordinarily because of a feeling of mild dyspnea. This feeling is greatly enhanced in persons who have a psychic fear of not being able to receive a sufficient quantity of air. For example, many persons on entering small or crowded rooms immediately experience emotional dyspnea. Also, patients with "cardiac neurosis" who have heard that dyspnea is associated with heart failure frequently experience severe psychic dyspnea. Neurogenic dyspnea has been known to be so intense that the person overrespires and causes alkalotic tetany.

HYPERCAPNIA

Hypercapnia means excess carbon dioxide in the body fluids.

One might suspect on first thought that any respiratory condition that causes hypoxia would also cause hypercapnia. However, hypercapnia usually occurs in association with hypoxia only when the hypoxia is caused by *hypoventilation* or by *circulatory deficiency*. The reasons for this are the following:

Obviously, hypoxia caused by *too little oxygen in the air*, by *too little hemoglobin*, or by *poisoning of the oxidative enzymes* has to do only with the availability of oxygen or use of oxygen by the tissues. Therefore, it is readily understandable that hypercapnia is *not* a concomitant of these types of hypoxia.

Also, in hypoxia resulting from poor diffusion through the pulmonary membrane or through the tissues, serious hypercapnia usually does not occur because carbon dioxide diffuses 20 times as rapidly as oxygen. Also, if hypercapnia does begin to occur this immediately stimulates pulmonary ventilation, which corrects the hypercapnia but not necessarily the hypoxia.

However, in hypoxia caused by hypoventilation, carbon dioxide transfer between the alveoli and the atmosphere is affected as much as is oxygen transfer. Therefore, hypercapnia always results along with the hypoxia. And in circulatory deficiency, diminished flow of blood decreases the removal of carbon dioxide from the tissues, resulting in tissue hypercapnia. However, the transport capacity of the blood for carbon dioxide is about three times that for oxygen, so that even here the tissue hypercapnia is much less than the tissue hypoxia.

EFFECTS OF HYPERCAPNIA ON THE BODY

When the alveolar P_{CO_2} rises above approximately 60 to 75 mm Hg, dyspnea usually becomes very severe, and as the P_{CO_2} rises to 80 to 100 mm Hg, the person becomes lethargic and sometimes even semicomatose. Anesthesia and death can result when the P_{CO_2} rises to 100 to 150 mm Hg.

OXYGEN THERAPY IN THE DIFFERENT TYPES OF HYPOXIA

Oxygen can be administered by (1) placing the patient's head in a "tent" that contains air fortified with oxygen, (2) allowing the patient to breathe either pure oxygen or high concentrations of oxygen from a mask, or (3) administering oxygen through an intranasal tube.

Oxygen therapy is of great value in certain types of hypoxia but of almost no value at all in other types. However, recalling the basic physiological principles of the different types of hypoxia, one can readily decide when oxygen therapy is of value and, if so, how valuable. For instance:

(1) In *atmospheric hypoxia*, oxygen therapy can obviously completely correct the depressed oxygen level in the inspired gases and, therefore, provide 100 per cent effective therapy.

(2) In *hypoventilation hypoxia*, a person breathing 100 per cent oxygen can move five times as much oxygen into the alveoli with each breath as when breathing normal air. Therefore, here again oxygen therapy can be extremely beneficial, increasing the available oxygen as much as fivefold. (However, this provides no benefit for the hypercapnia also caused by the hypoventilation.)

(3) In *hypoxia caused by impaired diffusion*, essentially the same result occurs as in hypoventilation hypoxia, for oxygen therapy can increase the P_{O_2} in the lungs from a normal value of about 100 mm Hg to as high as 600 mm Hg. This causes a greatly increased oxygen diffusion gradient between the alveoli and the blood, the gradient rising from a normal value of 60 mm Hg to as high as 560 mm Hg, or an increase of more than 800 per cent. This highly beneficial effect of oxygen therapy in diffusion hypoxia is illustrated in Figure 43–6, which shows that the pulmonary blood in this patient

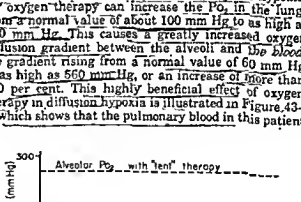


Figure 43–6. Absorption of oxygen into blood in pulmonary edema with and without ox

with pulmonary edema picks up oxygen four times as rapidly as it would with no therapy.

In *hypoxia caused by anemia or other abnormality of hemoglobin transport*, oxygen therapy is still of some value because a small amount of extra oxygen can be transported in the dissolved form in the fluids of the blood even though the amount of oxygen transported by the hemoglobin is hardly altered. This is illustrated in Figure 43–7, which shows that an increase in alveolar Po_2 from a normal value of 100 mm Hg to 600 mm Hg increases the total oxygen in the blood (that combined with hemoglobin plus the dissolved oxygen) from 5 volumes per cent to 6.5 volumes per cent. This represents a 30 per cent increase in the amount of oxygen transported to the tissues, and 30 per cent is often the difference between life and death. However, Figure 43–7 also shows that oxygen therapy in a *hyperbaric pressure chamber* at pressures above atmospheric pressure (760 mm Hg) can be of even greater value.

In *hypoxia caused by circulatory deficiency*, also called *ischemic hypoxia*, the value of oxygen therapy is usually very slight, because the problem here is sluggish flow of blood and not insufficient oxygen. For instance, Figure 43–8 illustrates that normal blood can carry only a small amount of extra oxygen to the tissues (about 10 per cent extra) when the alveolar oxygen concentration is increased to 600 mm Hg. Yet this 10 per cent difference may save the life of the patient, as, for example, following an acute heart attack which causes the cardiac output to fall very low.

In hypoxia caused by *physiologic shunt*, the value of oxygen therapy is even less than in ischemic hypoxia. Only that blood passing through the ventilated alveoli can receive additional oxygen. And since the blood passing through these alveoli is already either fully or almost fully saturated, the amount of additional oxygen that can be transported is only the portion that can become dissolved in the fluid of the nonshunted blood, a very small amount. Yet again, this may be the difference between life and death.

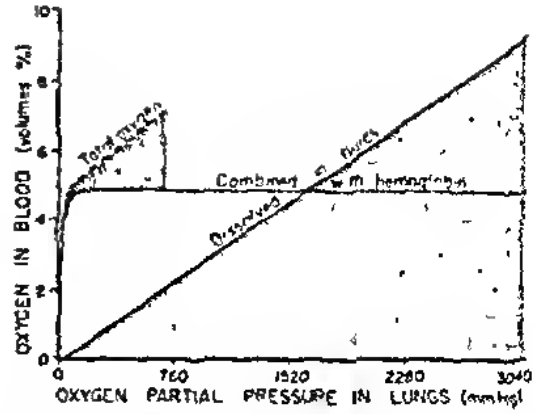


Figure 43–7. Effects of oxygen therapy in hypohemoglobinemic hypoxia.

In the different types of *hypoxia caused by inadequate tissue use of oxygen*, there is no abnormality of oxygen pickup by the lungs or of transport to the tissues. Instead, the tissue enzymes simply are not available to utilize the oxygen that is delivered. Therefore, it is doubtful that oxygen therapy is of any benefit at all.

DANGER OF HYPERCAPNIA DURING OXYGEN THERAPY

In the preceding chapter it was noted that in chronic hypoxia, oxygen lack becomes a far more powerful stimulus to respiration than usual, sometimes increasing the ventilation as much as five- to sevenfold. Therefore, during oxygen therapy, relief of the hypoxia occasionally causes the level of pulmonary ventilation to decrease so low that lethal levels of hypercapnia develop. For this reason, oxygen therapy in hypoxia is sometimes contraindicated, particularly in conditions that otherwise tend to cause hypercapnia, such as depressed respiratory center activity or airway obstruction.

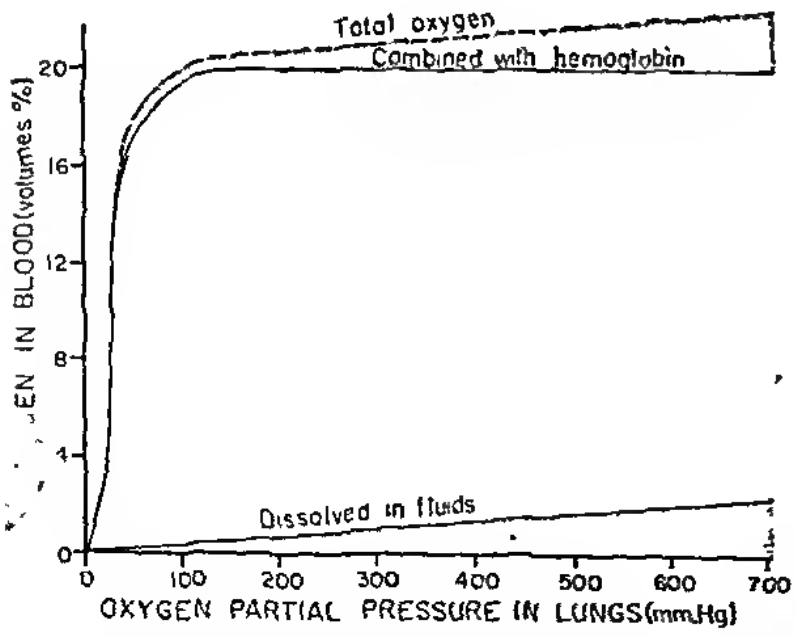


Figure 43–8. Effects of oxygen therapy in ischemic hypoxia.

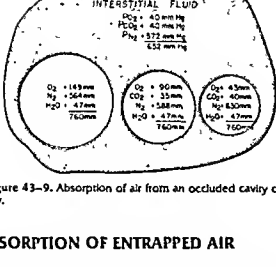


Figure 43–9. Absorption of air from an occluded cavity of the body.

ABSORPTION OF ENTRAPPED AIR

Whenever air becomes entrapped and closed off from the atmosphere anywhere in the body, whether in the alveoli as a result of bronchial obstruction, in the gastrointestinal tract, in one of the nasal or auditory cavities, or simply air injected beneath the skin, it is usually absorbed in a few hours or days.

Physical Principles of Air Absorption. Figure 43–9 shows the progressive stages of air absorption from a cavity. The cavity to the left is filled with recently injected air that has become humidified almost instantaneously by the fluids surrounding the cavity. Therefore, this cavity contains the normal concentrations of oxygen, nitrogen, and water vapor found in humidified air at sea level. Note that the total pressure of these three gases added together equals 760 mm Hg. Furthermore, the pressure in the cavity remains almost the same during the entire process of absorption because the atmospheric pressure continually presses against the body, and the body's tissues are flexible enough that the atmospheric pressure is transmitted almost exactly to most cavities beneath the skin.

At first, the PO_2 in the injected air is much greater than the PO_2 in the interstitial fluid, so that oxygen begins to be absorbed rapidly. On the other hand, the nitrogen pressure (PN_2) in the cavity is actually less than that in the interstitial fluid because nitrogen, unlike oxygen, is not metabolized by the tissues of the body. As a result, minute amounts of nitrogen actually diffuse into the cavity at first. But after large amounts of oxygen have been absorbed, the PO_2 falls to as little as 90 mm Hg, as shown in the second cavity in Figure 43–9. This absorption of oxygen makes the cavity become smaller, now increasing the PN_2 to a value greater than that in the interstitial fluid. Also, carbon dioxide diffuses into the cavity.

The final equilibrium state between the gases in the cavity and in the interstitial fluid is shown in the third cavity; the PO_2 has decreased to about 43 mm Hg, the PCO_2 has increased to 40 mm Hg, and the PN_2 has increased to 630 mm Hg. From then on, both oxygen and nitrogen are absorbed continually because both of these are then at higher pressures in the cavity than in the surrounding tissues. As this reduces the size of the cavity, the carbon dioxide pressure also increases to slightly higher than the pressure in the surrounding fluids. As a consequence, the carbon dioxide also is absorbed. And the water vapor returns to the liquid state. This process continues until all the gases leave the cavity and the cavity collapses totally.

REFERENCES

Busse, W W The precipitation of asthma by upper respiratory infections. *Chest*, 87 44S, 1985

Dawson, C A Role of pulmonary vasomotion in physiology of the lung *Physiol Rev*, *64* 544, 1984

Dempsey, J A., and Forster, H V Mediation of ventilatory adaptations *Physiol Rev*, 62 262, 1982.

Fanta, C H Calcium-channel blockers in prophylaxis and treatment of asthma *Am J Cardiol*, *55* 202B, 1985

Fisher, A B, *et al* Oxygen toxicity of the lung Biochemical aspects *In* Fishman, A P, and Renkin, E M (eds) Pulmonary Edema Baltimore, Waverly Press, 1979, p 207

Fishman, A P Assessment of Pulmonary Function New York, Mc Graw-Hill, 1980

Fishman, A P, and Pietra, C C Primary pulmonary hypertension *Annu Rev Med*, *31* 421, 1980

Guenter, C A, *et al.* Clinical Aspects of Respiratory Physiology Philadelphia, J B Lippincott Co, 1975

Guyton, A C, and Farish, C A A rapidly responding continuous oxygen consumption recorder *J Appl Physiol*, *14* 143, 1959

Hall, W J, and Douglas, R G, Jr Pulmonary function during and after common respiratory infections *Annu Rev Med*. *31* 233, 1980

Hodgkin, J E (ed) Chronic Obstructive Pulmonary Disease Current Concepts in Diagnosis and Comprehensive Care Park Ridge, Ill, American College of Chest Physicians, 1979

Kelsey, N Respiratory Therapy Review A Workbook and Study Guide St Louis, C V Mosby Co, 1982

Killian, K J, and Campbell, E J M Dyspnea and exercise *Annu. Rev Physiol*, *45* 465, 1983

Killough, J H Protective mechanisms of the lungs, pulmonary disease, pleural disease *In* Sodeman, W A, Jr and Sodeman, T M (eds). Sodeman's Pathologic Physiology, Mechanisms of Disease, 6th Ed Philadelphia, W B Saunders Co, 1979, p 451

Lane, D J Asthma. The Facts New York, Oxford University Press, 1979

Macklem, P T, and Permutt, S. (eds) The Lung in the Transition Between Health and Disease New York, Marcel Dekker, 1978

Malik, A B Pulmonary microembolism. *Physiol Rev*, 63 1114, 1983

Masuyama, H, and Honda, Y Differences in overall 'gain' of CO_2-feedback system between dead space and CO_2 ventilations in man. *Bull Eur Physiopathol Respir*, 20 501, 1984

McFadden, E R., Jr, and Ingram, R H, Jr Exercise-induced airway obstruction *Annu Rev Physiol*, *45* 453, 1983

Morris, H G Mechanisms of action and therapeutic role of corticosteroids in asthma. *J Allergy Clin Immunol*, *75* 1, 1985

Moser, K M (ed.) Pulmonary Vascular Diseases. New York, Marcel Dekker, 1979

Notter, R. H, and Finkelstein, J N Pulmonary surfactant. An interdisciplinary approach *J Appl Physiol*, *57* 1613, 1984

Packer, M Therapeutic application of calcium-channel antagonists for pulmonary hypertension *Am J Cardiol*, *55* 196B, 1985

Sant'Ambrogio, G Information arising from the tracheobronchial tree of mammals *Physiol Rev*, 62 531, 1982

Spector, S L The use of corticosteroids in the treatment of asthma Chest, *87* 73S, 1985

Staub, N C Pulmonary edema *Physiol Rev*, *54* 678, 1974

Stewart, R I Exercise in patients with chronic obstructive pulmonary disease. *S Afr Med J*, *67* 87, 1985

Way, J. L. Cyanide intoxication and its mechanism of antagonism *Annu. Rev Pharmacol Toxicol*, *24* 451, 1984

Wolfe, W. G, and Sabiston, D C Pulmonary Embolism. Philadelphia, W. B. Saunders Co, 1980

VIII

AVIATION, SPACE, AND DEEP SEA DIVING PHYSIOLOGY

44 ■ Aviation, High Altitude, and Space Physiology

45 ■ Physiology of Deep Sea Diving and Other Hyperbaric Conditions

44

Aviation, High Altitude, and Space Physiology

As man has ascended to higher and higher altitudes in aviation, in mountain climbing, and in space vehicles, it has become progressively more important to understand the effects of altitude and low gas pressures on the human body In the early days of aviation only two factors were of concern. (1) the effects of hypoxia on the body and (2) the effects of physical factors of high altitude, such as temperature and ultraviolet radiation With further development of airplanes, it was soon learned that they could be built to withstand accelerator forces far greater than the human body can stand Now, in the space age, all these problems have multiplied to the point where the physical conditions in the spacecraft must be created artificially

The present chapter deals with all these problems first, the hypoxia at high altitudes, second, some of the other physical factors affecting the body at high altitudes, and, third, the tremendous acceleratory forces that occur in both aviation and space physiology

EFFECTS OF LOW OXYGEN PRESSURE ON THE BODY

Barometric Pressures at Different Altitudes. As a prelude to discussing the effects of low oxygen pressure on the body, we must recall that the total pressure of all the gases in the air, the *barometric pressure*, decreases as one rises to progressively higher altitudes Table 44–1 gives the pressures at different altitudes, showing that at sea level the pressure is 760 mm Hg whereas at 10,000 feet it is only 523 mm Hg, and a 50,000 feet, 87 mm Hg This decrease in barometric pressure is the basic cause of all the hypoxia problem in high altitude physiology, for as the barometric pressure decreases, the oxygen pressure decreases proportionately, remaining at all times slightly less than 2 per cent of the total barometric pressure

Oxygen Partial Pressures at Different Elevations Table 44–1 also shows that the partial pressure o oxygen (Po_2) in dry air at sea level is approximately 15 mm Hg, though this can be decreased as much as 1 mm when large amounts of water vapor exist in the air The Po_2 at 10,000 feet is approximately 110 mm Hg; a 20,000 feet, 73 mm Hg; and at 50,000 feet, only 18 mm Hg

ALVEOLAR Po_2 AT DIFFERENT ELEVATIONS

Obviously, when the Po_2 in the atmosphere decrease at higher elevations, a decrease in alveolar Po_2 is als to be expected At low altitudes the alveolar Po_2 doe not decrease quite as much as the atmospheric Po because increased pulmonary ventilation helps compen sate for the diminished atmospheric oxygen But a higher altitudes the alveolar Po_2 decreases even mor than atmospheric Po_2, for peculiar reasons that an explained as follows:

Effect of Carbon Dioxide and Water Vapor o Alveolar Oxygen. Even at high altitudes carbon diox

Table 44–1 EFFECTS OF ACUTE EXPOSURE TO LOW ATMOSPHERIC PRESSURES ON ALVEOLAR GAS CONCENTRATIONS AND ON ARTERIAL OXYGEN SATURATION

			Breathing Air			Breathing Pure Oxygen		
Altitude (ft)	Barometric Pressure (mm Hg)	Po_2 in Air (mm Hg)	Pco_2 in Alveoli (mm Hg)	Po_2 in Alveoli (mm Hg)	Arterial Oxygen Saturation (%)	Pco_2 in Alveoli (mm Hg)	Po_2 in Alveoli (mm Hg)	Arterial Oxygen Saturatio (%)
0	760	159	40	104	97	40	673	100
10,000	523	110	36	67	90	40	436	100
20 000	349	73	24	40	73	40	262	100
30 000	226	47	24	21	30	40	139	99
40 000	141	29	24	12	15	36	58	87
50 000	87	18	24	2	2	24	16	22

de is continually excreted from the pulmonary blood nto the alveoli. Also, water vaporizes into the inspired air from the respiratory surfaces. Therefore, these two gases dilute the oxygen and nitrogen already in the alveoli, thus reducing the oxygen concentration.

The presence of carbon dioxide and water vapor in the alveoli becomes exceedingly important at high altitudes because the total barometric pressure falls to low levels while the pressures of carbon dioxide and water vapor do not fall comparably. Water vapor pressure remains 47 mm Hg as long as the body temperature is normal, regardless of altitude; and during acute exposure to very high altitudes the pressure of carbon dioxide falls from the normal sea level value of about 40 mm Hg to about 24 mm Hg because of increased respiration.

Now let us see how the pressures of these two gases affect the alveolar oxygen. For instance, assume that the total barometric pressure falls to 100 mm Hg; 47 mm Hg of this must be water vapor, leaving only 53 mm Hg for all the other gases. Under acute exposure to high altitude, 24 mm Hg of the 53 mm Hg must be carbon dioxide, leaving only 29 mm Hg. If there were no use of oxygen by the body, one fifth of this 29 mm Hg would be oxygen and four fifths would be nitrogen; or, the Po_2 in the alveoli would be 6 mm Hg. However, the person's tissues would be almost totally anoxic so that even most of this last remaining alveolar oxygen would be absorbed into the blood, leaving not more than 2 mm Hg oxygen pressure in the alveoli. Therefore, at a barometric pressure of 100 mm Hg, the person could not possibly survive when breathing air. But the effect is very different if the person is breathing pure oxygen, as we shall see in the following discussions.

A simple formula for approximating alveolar Po_2 is the following:

$$\text{Alveolar } Po_2 = \frac{Pb - Pco_2 - 47}{5} - Po_2 \text{ ABSORPTION}$$

In this formula, Pb is barometric pressure, the value 47 is the vapor pressure of water, and Po_2 ABSORPTION is the oxygen pressure decrease caused by oxygen uptake into the blood.

Alveolar Po_2 at Different Altitudes. The fifth column of Table 44–1 shows the Po_2s in the alveoli at different altitudes when one is breathing air. The alveolar Po_2 is 104 mm Hg at sea level; it falls to approximately 67 mm Hg at 10,000 feet and to only 2 mm Hg at 50,000 feet.

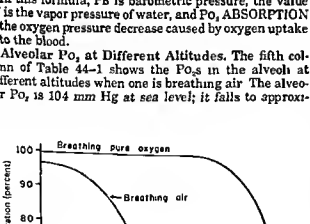


Figure 44–1. Effect of low atmospheric pressure on arterial oxygen saturation when breathing air and when breathing pure oxygen.

Saturation of Hemoglobin with Oxygen at Different Altitudes. Figure 44–1 illustrates arterial oxygen saturation at different altitudes when breathing air and when breathing oxygen, and the actual per cent saturation at each 10,000 foot level is given in Table 44–1. Up to an altitude of approximately 10,000 feet, even when air is breathed the arterial oxygen saturation remains at least as high as 90 per cent. However, above 10,000 feet the arterial oxygen saturation falls progressively, as illustrated by the left-hand curve of the figure, until it is only 70 per cent at 20,000 feet altitude and still less at higher altitudes.

EFFECT OF BREATHING PURE OXYGEN ON THE ALVEOLAR Po_2 AT DIFFERENT ALTITUDES

Referring once again to Table 44–1, note that when a person breathes air at 30,000 feet, the alveolar Po_2 is only 21 mm Hg even though the barometric pressure is 226 mm Hg. The most important cause of this difference is that almost three fourths of the alveolar air is nitrogen. But if the person breathes pure oxygen instead of air, most of the space in the alveoli formerly occupied by nitrogen now becomes occupied by oxygen instead. Theoretically, at 30,000 feet the aviator could have an alveolar Po_2 of 139 mm Hg instead of the 21 mm Hg when breathing air.

The second curve of Figure 44–1 illustrates the arterial oxygen saturation at different altitudes when breathing pure oxygen. Note that the saturation remains above 90 per cent until the aviator ascends to approximately 39,000 feet; then it falls rapidly to approximately 50 per cent at about 47,000 feet.

THE "CEILING" WHEN BREATHING AIR AND WHEN BREATHING OXYGEN IN AN UNPRESSURIZED AIRPLANE

Comparing the two arterial oxygen saturation curves in Figure 44–1, one notes that an aviator breathing oxygen can ascend to far higher altitudes than one not breathing oxygen. For instance, the arterial saturation at 47,000 feet when breathing oxygen is about 50 per cent and is equivalent to the arterial oxygen saturation at 23,000 feet when breathing air. And, because a person ordinarily can remain conscious until the arterial oxygen saturation falls to 40 to 50 per cent, the ceiling for an unacclimatized aviator in an unpressurized airplane when breathing air is approximately 23,000 feet and when breathing pure oxygen about 47,000 feet, provided the oxygen-supplying equipment operates perfectly.

EFFECTS OF HYPOXIA

The rate of pulmonary ventilation ordinarily does not increase significantly until one has ascended to about 8000 feet. At this height the arterial oxygen saturation has fallen to approximately 93 per cent, at which level the chemoreceptors begin to respond significantly. Above 8000 feet the chemoreceptor stimulatory mechanism progressively increases the ventilation until one reaches approximately 16,000 to 20,000 feet, at which altitude the ventilation has reached a maximum about 65 per

cent above normal if the person is exposed only acutely to the high altitude. (But several days of exposure increases ventilation about 400 per cent.)

Other effects of hypoxia, beginning at an altitude of approximately 12,000 feet, are drowsiness, lassitude, mental fatigue, sometimes headache, occasionally nausea, and sometimes euphoria. Most of these symptoms increase in intensity at still higher altitudes, the headache often becoming especially prominent and the cerebral symptoms sometimes progressing to the stage of twitchings or convulsions and ending, above 23,000 feet in the unacclimatized person, in coma.

One of the most important effects of hypoxia is decreased mental proficiency, which decreases judgment, memory, and the performance of discrete motor movements. Ordinarily these abilities remain absolutely normal up to approximately 9000 feet, and may be completely normal for a short time up to elevations of 15,000 feet. But, if the aviator is exposed to hypoxia for a long time, mental proficiency, as measured by reaction times, handwriting, and psychological tests, may decrease to 80 per cent of normal even at altitudes as low as 11,000 feet. If an aviator stays at 15,000 feet for one hour, mental proficiency ordinarily will have fallen to approximately 50 per cent of normal, and after 18 hours at this level to approximately 20 per cent of normal.

Respiratory Center Depression as the Cause of Death in Hypoxia. When a person develops severe hypoxia—that is, to the stage of coma—the respiratory center itself often becomes depressed within another few minutes because of metabolic deficit in its neuronal cells. This counteracts the stimulatory effect of the chemoreceptor mechanism, and the respiration, instead of increasing further, decreases precipitously until it actually ceases.

Effects of Sudden Exposure to Low Po_2. At times the aviator who has been flying at a very high altitude with special oxygen equipment or in a pressurized cabin suddenly becomes detached from the oxygen supply, or the cabin decompresses. The alveolar Po_2 falls within a few seconds to a low value, but because oxygen is stored in the body fluids (combined with hemoglobin and with various oxygen carriers of the tissues), a short time elapses before the body suffers dire results from oxygen lack.

Figure 44–2 illustrates graphically the time that ordinarily elapses, first, before the aviator shows signs of diminished consciousness, and, second, before actual coma results. Note from this figure that at 38,000 feet diminished consciousness begins in approximately 30 seconds, and coma results in approximately one minute, while at 28,000 feet diminished consciousness begins in approximately one minute, and coma occurs in approximately three minutes *in the unacclimatized person*.

ACCLIMATIZATION TO LOW Po_2

A person remaining at high altitudes for days, weeks, or years gradually becomes acclimatized to the low Po_2 so that it causes fewer and fewer deleterious effects on the body and also so that it becomes possible for the person to work harder without hypoxic effects or to ascend to still higher altitudes. The five principal means by which acclimatization comes about are (1) increased pulmonary ventilation, (2) increased hemoglobin in the blood, (3) increased diffusing capacity of the lungs, (4) increased vascularity of the tissues, and (5) increased ability of the cells to utilize oxygen despite the low Po_2.

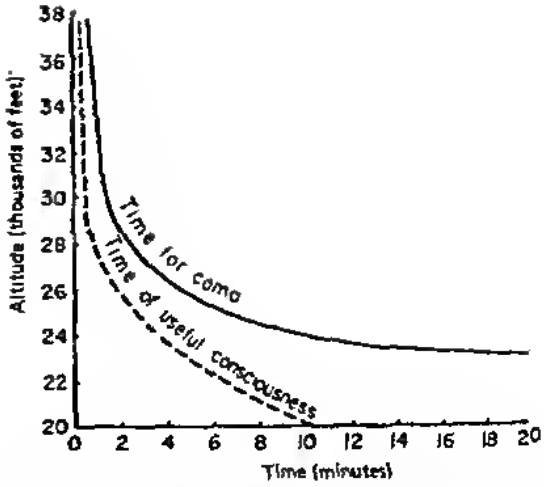


Figure 44–2. Time of exposure to low oxygen concentrations required to cause diminished consciousness or coma. (From Armstrong, Principles and Practice of Aviation Medicine. The Williams & Wilkins Co.)

Increased Pulmonary Ventilation. On immediate exposure to low Po_2, the hypoxic stimulation of the chemoreceptors increases alveolar ventilation to a maximum of about 65 per cent. This is an immediate compensation for the high altitude, and it alone allows the person to rise several thousand feet higher than would be possible without the increased ventilation. Then, if the person remains at a very high altitude for several days, the ventilation gradually increases to as much as three to seven times normal. The basic cause of this gradual increase is the following:

The immediate 65 per cent increase in pulmonary ventilation on rising to a high altitude blows off large quantities of carbon dioxide, reducing the Pco_2 and increasing the pH of the body fluids. Both these changes *inhibit* the respiratory center and thereby *oppose the stimulation by the hypoxia*. However, during the ensuing two to five days, this inhibition fades away, allowing the respiratory center now to respond with full force to the chemoreceptor stimuli resulting from hypoxia, and the ventilation increases to about three to seven times normal. The cause of this fading inhibition is unknown, but there is some evidence that it results from reduced CSF bicarbonate ion and perhaps also reduced brain tissue bicarbonate ion. These changes, in turn, decrease the pH in the fluids surrounding the chemosensitive neurons of the respiratory center, thus increasing the activity of the center. (See Chapter 42.)

Rise in Hemoglobin During Acclimatization. It will be recalled from Chapter 4 that hypoxia is the principal stimulus for causing an increase in red blood cell production. Ordinarily, in full acclimatization to low oxygen the hematocrit rises from a normal value of 40 to 45 to an average of 60 to 65, with an average increase in hemoglobin concentration from a normal of 15 gm per cent to about 22 gm per cent.

In addition, the blood volume also increases, often by as much as 20 to 30 per cent, resulting in a total increase in circulating hemoglobin of as much as 50 to 90 per cent.

Unfortunately, this increase in hemoglobin and blood volume is a slow one, having almost no effect until after two to three weeks, reaching half development in a

ıonth or so, and becoming fully developed only after ıany months.

Decreased Affinity of the Hemoglobin for Oxygen ʼnder Hypoxic Conditions. Within a few hours after ıe blood is first exposed to hypoxia at high altitudes, ıcreased quantities of phosphate compounds are formed ıside the red blood cells, and some of these combine ıith the hemoglobin to decrease its affinity for oxygen, ıat is, to shift the hemoglobin-oxygen saturation curve ıward higher Po_2s, as was illustrated in Chapter 41. ıne of these, 2,3-diphosphoglycerate, commonly called ,3-DPG, is especially significant. Because of the reduced ffinity for oxygen, tbe hemoglobin delivers the oxygen ı the tissue cells at a higher Po_2. At altitudes up to bout 15,000 feet this effect increases oxygen delivery s much as 10 to 20 per cent. But at still higher altitudes ıe decreased affinity for oxygen also decreases the of oxygen by the hemoglobin in the lungs, and ıerefore it decreases the overall availability of oxygen) the tissues, which can actually become more harmful ıan the increased tendency for oxygen release by the emoglobin at the tissue level is helpful.

Increased Diffusing Capacity During Acclimatization. It will be recalled that the normal diffusing apacity for oxygen through the pulmonary membrane ı approximately 21 ml per mm Hg pressure gradient er minute, and this diffusing capacity can increase as uch as threefold during exercise. A similar increase n diffusing capacity occurs at high altitude Part of the ncrease probably results from greatly increased pulnonary capillary blood volume, which expands tha capllaries and increases the surface through which oxygen an diffuse into the blood. Another part results from an ncrease in lung volume, which expands the surface area f the alveolar membrane. A final part results from an ncrease in pulmonary arterial pressure; this forces lood into greater numbers of alveolar capillaries than ıormally—especially in the upper parts of the lungs, which are poorly perfused under usual conditions.

The Circulatory System in Acclimetization—Increased Vascularity. The cardiac output often increases as much as 20 to 30 per cent immediately after ı person ascends to a high altitude, but it usually falls ıack to normal within a few days. Then, over a period f months, the cardiac output decreases a moderate ımount, but this is associated with e simultaneous *ncrease in blood hematocrit* so that the amount of ıxygen transported to the tissues remains approximately normal—that is, unless the altitude becomes so ıigh that severe hypoxia results.

Another circulatory adaptation is an increase in numbers and sizes of capillaries in the tissues, which is called *increased vascularity.* This occurs especially in animals born and bred at high altitudes, but less so for animals that later in life have become exposed to the altitude. In very active tissues exposed to chronic hypoxia, the increase in vascularity is especially marked. For instance, the capillary density in the right ventricular *muscle increases many per cent because of the* combined effects of hypoxia and excess work load on the right ventricle, caused by pulmonary hypertension at high altitude (the hypertension results from pulmonary vasoconstriction caused by low alveolar oxygen concentration).

Cellular Acclimatization. In animals native to altitudes of 13,000 to 17,000 feet, mitochondria and certain cellular oxidative enzyme systems are more plentiful than in sea level inhabitants. Therefore, it is presumed that acclimatized human beings as well as these animals can utilize oxygen more effectively than can their sea level counterparts.

NATURAL ACCLIMATIZATION OF NATIVES LIVING AT HIGH ALTITUDES

Many natives in the Andes and in the Himalayas live at altitudes above 13,000 feet—one group in the Peruvian Andes actually lives at an altitude of 17,500 feet and works a mine at an altitude of 19,000 feet. Many of these natives are born at these altitudes and live there all their lives In all the aspects of acclimatization just listed the natives are superior to even the best-acclimatized lowlanders, even though the lowlanders might have also lived at high altitudes for ten or more years. This process of acclimatization of the natives begins in infancy The chest size, especially, is greatly increased whereas the body size is somewhat decreased, giving a high ratio of ventilatory capacity to body mass In addition, their hearts, particularly the right heart which provides a high pulmonary arterial pressure to pump blood through a greatly expanded pulmonary capillary system, are considerably larger than the hearts of lowlanders

The delivery of oxygen by the blood to the tissues is also highly facilitated in these natives. For instance, Figure 44–3 shows the hemoglobin-oxygen dissociation curves for natives who liva at sea level and for their counterparts who live at 15,000 feet Note that the arterial oxygen Po_2 in the natives at higher altitude is only 40 mm Hg, but because of the greater quantity of hemoglobin the quantity of oxygen in the arterial blood is actually greater than in the blood of the natives at the lower altitude. Note also that the venous Po_2 in the high altitude natives is only 10 mm Hg less than the venous Po_2 for the lowlanders, despite the low arterial Po_2, indicating that oxygen transport to the tissues is exceedingly effective in the naturally acclimatized high altitude natives.

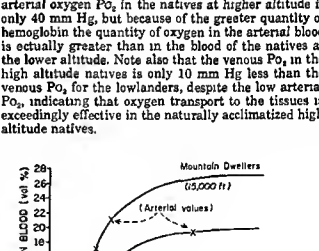


Figure 44–3. Oxygen-dissociation curves for blood of high-altitude and sea-level residents, showing the respective arterial and venous Po_2s and oxygen contents as recorded in their native surroundings. (From "Oxygen-dissociation curves for bloods of high-altitude and sea-level residents." *PAHO Scientific Publication* No. 140, Life at High Altitudes, 1966.)

WORK CAPACITY AT HIGH ALTITUDES: THE EFFECT OF ACCLIMATIZATION

Aside from the mental depression caused by hypoxia, as discussed earlier, the work capacity of all the muscles is also greatly decreased in hypoxia This includes not only the skeletal muscles but also the cardiac muscle so that even the maximum level of cardiac output is reduced. In general, the work capacity is reduced in direct proportion to the decrease in maximum rate of oxygen uptake that the body can achieve

To give an idea of the importance of acclimatization for work capacity, consider this The work capacities in per cent of sea level maximum for unacclimatized and acclimatized persons at an altitude of 17,000 feet have been measured to be the following

	per cent
Unacclimatized	50
Acclimatized for two months	68
Native living at 13,200 feet but working at 17,000 feet	87

Thus, naturally acclimatized natives can achieve a daily work output even at these high altitudes almost equal to that of a normal person at sea level, but even well-acclimatized lowlanders can almost never achieve this result.

To emphasize further the importance of natural acclimatization of natives, the Sherpas of the Himalayas can survive without oxygen for many minutes and perhaps even hours at altitudes as high as Mt Everest—over 29,000 feet!

CHRONIC MOUNTAIN SICKNESS

Occasionally, a person who remains at high altitude too long develops chronic mountain sickness, in which the following effects occur: (1) the red cell mass and hematocrit become exceptionally high, (2) the pulmonary arterial pressure becomes elevated even more than the normal elevation that occurs during acclimatization, (3) the right heart becomes greatly enlarged, (4) the peripheral arterial pressure begins to fall, (5) congestive failure ensues, and (6) death often follows unless the person is removed to a lower altitude The cause of this sequence of events is probably threefold First, the red cell mass becomes so great that the blood viscosity increases severalfold, this now actually *decreases* tissue blood flow so that oxygen delivery begins to decrease as well Second, the pulmonary arterioles become vasospastic because of the lung hypoxia This results from the hypoxic vascular constrictor effect that normally operates to divert blood flow from unaerated to aerated alveoli, as explained in Chapter 24. But in this instance all the arterioles become constricted, the pulmonary arterial pressure rises excessively, and the right heart fails Third, the pulmonary arteriolar spasm diverts much of the blood flow through nonalveolar pulmonary vessels, thus causing an excess of pulmonary shunt blood flow where the blood is not oxygenated; and this further compounds the problem Fortunately, most of these persons recover within days or weeks when they are moved to lower altitude.

ACUTE MOUNTAIN SICKNESS AND HIGH ALTITUDE PULMONARY EDEMA

A small per cent of persons who ascend rapidly to high altitude become acutely sick and can die if not given oxygen or removed to a low altitude. The si begins from a few hours up to about two days after ascent. Two different events frequently occur in th sickness

(1) *Acute cerebral edema* This is believed to from local vasodilatation of the cerebral blood vess caused by the hypoxia Dilatation of the arterioles i creases the capillary pressure, which in turn fluid to leak into the cerebral tissues The edema can then lead to severe disorientation and effects related to cerebral dysfunction

(2) *Acute pulmonary edema.* This is believed to caused in the following way The severe hypox the pulmonary arterioles to constrict potently, but constriction is much greater in some parts of the than in other parts so that more and more of pulmonary blood flow is forced through fewer and still unconstricted pulmonary vessels The result is the capillary pressure in these areas of the lungs be comes especially high and local edema occurs extension of the process to progressively more areas of the lungs leads to severe pulmonary dysfuncti that can be lethal However, allowing the person breathe oxygen will usually reverse the process withi hours

Some persons have extreme pulmonary vascular rea tivity to hypoxia, many times the reactivity of th normal person; these people are especially susceptibl to acute high altitude pulmonary edema.

EFFECTS OF ACCELERATORY FORCES ON THE BODY IN AVIATION AND SPACE PHYSIOLOGY

Because of rapid changes in velocity and direction motion in airplanes and space ships, several types acceleratory forces often affect the body during fligh At the beginning of flight, simple linear acceleratio occurs; at the end of flight, deceleration; and, every tim the vehicle turns, angular and centrifugal acceleratio In aviation physiology it is usually centrifugal accele ation that demands greatest consideration because th structure of the airplane is capable of withstandin much greater centrifugal acceleration than is the huma body

CENTRIFUGAL ACCELERATORY FORCES

When an airplane makes a turn, the force of centri ugal acceleration is determined by the following rel tionship.

$$f = \frac{mv^2}{r}$$

in which f is the centrifugal acceleratory force, m is th mass of the object, v is the velocity of travel, and r the radius of curvature of the turn. From this formul it is obvious that as the velocity increases, the force centrifugal acceleration increases in proportion to th *square* of the velocity. It is also obvious that the for of acceleration is directly proportional to the sharpne of the turn (the less the radius)

Measurement of Acceleratory Force—"G." Whe a man is simply sitting in his seat, the force with whic he is pressing against the seat results from the pull gravity, and it is equal to his weight. The intensity

his force is said to be +1 "G" because it is equal to the full of gravity. If the force with which he presses against he seat becomes five times his normal weight during ull-out from a dive, the force acting upon the seat is -5 G.

If the airplane goes through an outside loop so that the man is held down by his seat belt, *negative* G is applied to his body, and, if the force with which he is thrown against his belt is equal to the weight of his body, the negative force is −1 G

Effects of Centrifugal Acceleratory Force on the Body. *Effects on the Circulatory System.* The most important effect of centrifugal acceleration is on the circulatory system because blood is mobile and can be translocated by centrifugal forces Centrifugal forces also tend to displace the tissues but, because of their solid structure, they only sag—ordinarily not enough to cause abnormal function.

When the aviator is subjected to *positive* G the blood is centrifuged toward the lower part of the body Thus, if the centrifugal acceleratory force is +5 G and the person is in an immobilized standing position, the hydrostatic pressure in the veins of the feet is five times normal, or approximately 450 mm Hg; even in the sitting position this pressure is nearly 300 mm Hg As the pressure in the vessels of the lower part of the body increases, the vessels passively dilate, and a major proportion of the blood from the upper part of the body is translocated into these lower vessels. Because the heart cannot pump unless blood returns to it, the greater the quantity of blood "pooled" in the lower body the less becomes the cardiac output

Figure 44–4 illustrates the changes in systolic and diastolic arterial pressures when an acceleratory force of +3.3 G is suddenly applied Note that both these pressures fall below 20 mm Hg for the first few seconds after the acceleration begins but then return to a systolic pressure of about 55 mm Hg and a diastolic pressure of 20 mm Hg within another 10 to 15 seconds This secondary recovery is caused mainly by *activation of the* baroreceptor reflexes

Acceleration greater than 4 to 6 G causes "blackout" of vision within a few seconds and unconsciousness shortly thereafter. If this degree of acceleration is continued, the person will die.

Effects on the Vertebrae. Extremely high acceleratory forces for even a fraction of a second can fracture the vertebrae. The degree of positive acceleration that the average person can withstand in the sitting position before vertebral fracture occurs is approximately 20 G

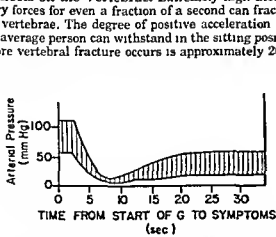


Figure 44–4. Changes in systolic and diastolic arterial pressures following abrupt and continuing exposure to an acceleratory force of +3.3 G. (Modified from Martin and Henry: *J. Aviation Med.* 22:382, 1951)

This force also approaches the limit of safety for the structural elements of most airplanes. Therefore, any attempt to utilize such intense positive G as this in maneuvers would be extremely dangerous for reasons other than simply the effects on the body

Transverse G. The human body can withstand tremendous transverse acceleratory forces (forces applied along the antero-posterior axis of the body—lying down in the airplane, for instance) If the transverse acceleratory forces are applied uniformly over large areas of the body, as much as 15 to 25 G can be withstood for many seconds without serious effects other than occasional collapse of a lung, which is not a lethal effect Therefore, when very large acceleratory forces are to be involved, the aviator or astronaut flies in the semi-reclining or lying position

Negative G. The effects of negative G on the body are less dramatic acutely but possibly more damaging permanently than the effects of positive G An aviator can usually go through outside loops up to negative acceleratory forces of −4 to −5 G without harm except for intense momentary hyperemia of the head, though occasionally psychotic disturbances lasting for 15 to 20 minutes occur thereafter as a result of brain edema

Occasionally negative G forces can be so great and centrifugation of the blood into the head so great that the cerebral blood pressure reaches 300 to 400 Hg, sometimes causing small vessels on the surface of the head and in the brain to rupture However, the vessels inside the cranium show less tendency for rupture than would be expected, for the cerebrospinal fluid is centrifuged toward the head at the same time that blood is centrifuged toward the cranial vessels, and the greatly increased pressure of this fluid acts as a cushioning buffer on the outside of the brain to prevent vascular rupture Even so, animals exposed to negative acceleratory forces of 20 to 40 G have developed subarachnoid hemorrhages, and undoubtedly such effects could occur in the human being.

Because the eyes are not protected by the cranium, intense hyperemia occurs in them during negative G As a result, the eyes often become temporarily blinded with "redout"

Protection of the Body Against Centrifugal Acceleratory Forces. Specific procedures and apparatus have been developed to protect aviators against the circulatory collapse that occurs during positive G First, if the aviator tightens the abdominal muscles to an extreme degree and leans forward to compress the abdomen, some of the pooling of blood in the large vessels of the abdomen can be prevented, thereby delaying the onset of blackout Also, special "anti-G" suits have been devised to prevent pooling of blood in the lower abdomen and legs. The simplest of these applies positive pressure to the legs and abdomen by inflating compression bags as the G increases Theoretically, a pilot submerged in a tank or suit of water could withstand very high degrees of acceleration, both positive and negative, for the pressures developed in the water and pressing on the outside of the body during centrifugal acceleration almost exactly balance the forces acting on the body. Unfortunately, however, the presence of air in the lungs still allows displacement of the heart, the lung tissues, and the diaphragm into seriously abnormal positions despite submersion in water. Therefore, even if this procedure were used, the limits of safety would be about 15 to 20 G.

EFFECTS OF LINEAR ACCELERATORY FORCES ON THE BODY

Acceleratory Forces in Space Travel. Unlike an airplane, a spacecraft cannot make rapid turns; therefore, centrifugal acceleration is of little importance except when the spacecraft goes into abnormal gyrations. On the other hand, blast-off acceleration and landing deceleration might be tremendous; both of these are types of linear acceleration

Figure 44–5 illustrates a typical profile of the acceleration during blast-off in a three-stage spacecraft, showing *that the first stage booster causes acceleration* as high as 9 G and the second stage booster, as high as 8G In the standing position the human body could not withstand this much acceleration, but in a semi-reclining position *transverse to the axis of acceleration*, this amount of acceleration can be withstood with ease despite the fact that the acceleratory forces continue for as long as five minutes at a time. Therefore, we see the reason for the reclining seats used by the astronauts.

Problems also occur during deceleration when the spacecraft re-enters the atmosphere. A person traveling at Mach 1 (the speed of sound and of fast airplanes) can be safely decelerated in a distance of approximately 0 12 mile, whereas a person traveling at a speed of Mach 100 (a speed possible in interplanetary space travel) requires a distance of about 10,000 miles for safe deceleration The principal ceason for this difference is that the total amount of energy that must ba dispelled during deceleration ia proportional to the *square* of the velocity, which alone increases the distance 10,000-fold But, in addition to this, a human being can withstand far less deceleration if the period of deceleration lasts for a long time than for a short time. Therefore, deceleration must be accomplished much more slowly from the very high velocitiea than is necessary at the slower velocities

Doceleratory Forces Aasociated with Parachute Jumps. When the parachuting aviator leaves the airplane, the velocity of fall is exactly 0 feet per second at first. However, because of the acceleratory force of gravity, within 1 aecond the velocity of fall is 32 feet per second (if there is no air resistance), in two seconds it is 64 feet per second, and so on However, as the rate of fall increases, the air resistance tending to slow the fall also increases Finally, the deceleratory force of the air resistance is equal to the acceleratory force of gravity so that after falling for approximately 12 seconds and a distance of 1400 feet, the person will be falling at a "terminal velocity" of 109 to 119 miles per hour (17[illegible] feet per second)

If the parachutist has already reached the terminal velocity of fall before opening the parachute, an "opening shock load" of approximately 1200 pounds occurs on the parachute shrouds

The usual-sized parachute slows the fall of the parachutist to approximately one ninth the terminal velocity In other words, the speed of landing is approximately 20 feet per second, and the force of impact against the earth is approximately 1/81 the impact force without a parachute Even so, the force of impact is still great enough to cause considerable damage to the body unless the parachutist is properly trained in landing. Actually the force of impact with the earth is approximately the same as that which would be experienced by jumping from a height of about 6 feet. Unless forewarned, the parachutist will be tricked by the senses into striking the earth with extended legs, and this will result in tremendous deceleratory forces along the skeletal axis of the body, resulting in fracture of the pelvis, of vertebrae, or of a leg Consequently, the trained parachutist strikes the earth with kneea bent, but muscles taut, to cushion the shock of landing.

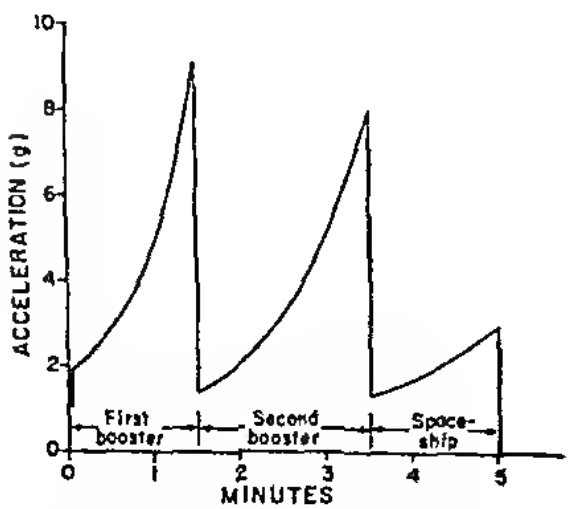


Figure 44–5. Acceleratory forces during the takeoff of a space-

PERCEPTIONS OF EQUILIBRIUM AND TURNING IN BLIND FLYING

The various aensations of equilibrium will be discussed in Chapter 52, but in essence, whenever the head is not positioned along the vertical axis of forces applied to the body, the otoliths in the utriclea and saccules of the vestibular apparatus apprise the psyche of a lack of equilibrium However, the degree of proficiency of equilibrium perceptions has certain limits For instance, the body must lean forward as much aa 5 to 10 degrees before the utricles sense this forward leaning. Therefore when the aviator is flying blind, it is necessary that appropriate instruments be available to indicate the rate of descent. Also, the vestibular apparatus may fail to apprise the pilot of ascent up to an error of as much as 24 degrees *Obviously, then, blind flying often results* in stall of the airplane unless appropriate instruments are available.

Perhaps the least effective of the perceptive organs in blind flying are those for turning, because the organs that perceive turning—the semicircular canals—are excited only for the first few seconds after a turn begins and only for the first few seconds after the turn ends Consequently, an aviator might perceive beginning to turn, but once in the turn, all perception of continuing the turn may be lost. Furthermore, if one goes into the turn slowly—at an angular acceleratory rate less than 2 degrees per second per second—the fact of entering the turn may not even be perceived Consequently, of especial importance among the instruments for blind flying are thosa that apprise the aviator of the direction of travel of the airplane and the rate of turn

PROBLEMS OF TEMPERATURE IN AVIATION AND SPACE PHYSIOLOGY

The cold temperatures of the upper atmosphere involve essentially the same physiological problems as

ıld temperatures on the surface of the earth; these ·oblems will be discussed in Chapter 72. Therefore, for ıe present it shall only be pointed out how the temper- ıtures change as one ascends to higher elevations. If ıe temperature at the earth's surface is approximately)° C, temperatures at different altitudes are approxi- ıately the following:

Feet	°C
0	20
10,000	0
20,000	−22
30,000	−44
40,000	−55

It is obvious from these values that special clothing r special heating apparatus must be designed for flying t high altitudes.

Temperature in Space. Though the temperature of he air falls to about −55° C several miles above the arth, the temperature rises to very high values in pace. In the ionosphere several hundred miles above he earth and in space, the kinetic energy of the few nolecules, atoms, and ions that are there is extreme. 'he reason for this is that any particles that can escape his far away from the gravitational pull of the celestial ·odies must have a great velocity. As a result, by the ime the spacecraft has reached an altitude above 350 ıiles the temperature of the surrounding particles is ıbout 3000° C. Yet, strangely enough, this has almost ıo effect on the temperature of the spacecraft because ıf the *sparsity* of these particles—that is, they are far oo few to impart any significant amount of heat to the spacecraft. Instead, the temperature of the spacecraft is letermined by the relative *absorption of radiant energy* rom the sun versus the *re-radiation of energy* away rom the spacecraft. Different coatings for the spacecraft ıave different absorptive and radiation characteristics. Therefore, with appropriate coating and orientation of he spacecraft appropriately with respect to the sun, the temperature inside can be made almost any desired value.

ADIATION AT HIGH ALTITUDES AND IN SPACE

Radiation in the Atmosphere. The different types of radiation change considerably between the surface of he earth and at high altitudes because the atmosphere filters some of the radiations from the sun and other celestial sources before they reach the earth's surface. Ordinarily most of the ultraviolet light is absorbed before it reaches the earth's surface. Consequently, a person exposed to the sun in the upper atmosphere is many times as likely to develop sunburn as on the earth's surface.

Approximately 18 per cent of the visible light is absorbed by the atmosphere before it reaches the surface of the earth even on a perfectly clear day. Therefore, the brightness of the sun is 1.2 times as great in the upper atmosphere as on earth. This makes the earth less distinct to the aviator than otherwise for two reasons: first, he or she is looking from an area of greater brightness toward an area of lesser brightness, and second, light is reflected from the atmosphere into his eyes. This reflected light especially blocks his vision of the horizon.

In aviation there is no significant hazard from gamma rays, x-rays, and cosmic rays, but this does become a problem in space physiology.

Radiation Hazards in Space Physiology. Large quantities of cosmic particles, especially protons and electrons, are continually bombarding the earth's upper atmosphere, some originating from the sun and some from outer space. The magnetic field of the earth traps many of these cosmic particles in two major belts around the earth called *Van Allen radiation belts*. The inner belt begins at an altitude of about 300 miles and extends to about 3000 miles. The outer belt begins at about 6000 miles and extends to 20,000 miles. However, because of the magnetic field of the earth, these belts occur mainly in the tropical and temperate zones around the earth and not at the two poles.

The types of radiation in the two Van Allen belts are almost entirely high energy electrons and protons. A person traversing these two belts during an interplanetary space trip can receive as much as 1 rad of radiation, which is about 1/400 the lethal dose, and a person in a spacecraft orbiting the earth within one of these two belts for prolonged periods of time could receive enough radiation to cause death.

Thus, it is usually important to orbit spacecraft below an altitude of 200 to 300 miles, an altitude at which the radiation hazard is slight.

"ARTIFICIAL CLIMATE" IN THE SEALED SPACECRAFT

Since there is no atmosphere in outer space, an artificial atmosphere and climate must be provided. Most important of all, the oxygen concentration must remain high enough and the carbon dioxide concentration low enough to prevent suffocation. In some of the space missions a capsule atmosphere containing pure oxygen at about 260 mm Hg pressure has been used. In others, a gas mixture of 50 per cent oxygen and 50 per cent nitrogen at a total pressure of about 380 mm Hg has been used. The presence of nitrogen in the mixture greatly diminishes the likelihood of fire and explosion.

For space travel lasting more than several months, it will be impractical to carry along an adequate oxygen supply and enough carbon dioxide absorbent. For this reason, "recycling techniques" have been developed for use of the same oxygen over and over again. These techniques also frequently include re-use of the same food and water. Basically, they involve (1) a method for removing oxygen from carbon dioxide, (2) a method for removing water from the human excreta, and (3) use of the human excreta for resynthesizing or regrowing an adequate food supply.

Large amounts of energy are required for these processes, and the real problem at present is to derive enough energy from the sun's radiation to energize the necessary chemical reactions. Some recycling processes depend on purely physical procedures, such as distillation, electrolysis of water, and capture of the sun's energy by solar batteries, whereas others depend on biological methods, such as use of algae with their large store of chlorophyll to generate foodstuffs by photosynthesis. Unfortunately, a completely practical system for recycling is yet to be achieved. The problem is the weight of the equipment that must be carried.

WEIGHTLESSNESS IN SPACE

A person in an orbiting satellite or in any nonpropelled spacecraft experiences *weightlessness* That is, the person is not drawn toward the bottom, sides, or top of the spacecraft but simply floats inside its chambers The cause of this is not failure of gravity to pull on the body, because gravity from any nearby heavenly body is still active However, the gravity acts on both the spacecraft and the person at the same time, and since there is no resistance to movement in space, both are pulled with exactly the same acceleratory forces and in the same direction. For this reason, the person simply is not attracted toward any wall of the spacecraft

Weightlessness is mainly an engineering problem—that is, to provide special techniques for eating and drinking (since food and water will not stay on open plates or in glasses), special waste disposal systems (especially for the excreta), and adequate hand holds or other means for stabilizing the person in the spacecraft for adequately controlling the operation of the ship

Physiological Problems of Weightlessness. Fortunately, the physiological problems of weightlessness have not proved to be severe Most of the problems that do occur appear to be related to two effects of the weightlessness. (1) translocation of fluids within the body because of failure of gravity to cause hydrostatic pressures, and (2) diminishment of physical activity because no strength of muscle contraction is required to oppose the force of gravity

The observed effects of prolonged stay in space are the following (1) decrease in blood volume, (2) decrease in red cell mass, (3) decreased muscle strength and work capacity, (4) decrease in maximum cardiac output, and (5) loss of calcium from the bones Essentially these same effects also occur in persons lying in bed for an extended period of time For this reason extensive exercise programs are carried out during prolonged Space Laboratory missions, and all the above effects are greatly reduced In previous Space Lab expeditions, in which the exercise program had been less vigorous, the astronauts had considerably lower work capacities for the first few days after returning to earth They also had a tendency to faint when they stood up during the first day or so after return to gravity because of translocation of their diminished blood volume toward their abdomens and legs

The effects of weightlessness that do occur usually reach their maximum within the first few weeks after the astronaut enters the space environment and fortunately are not progressive thereafter with one exception the calcium loss continues for many months, presumably because the stimulation of bone deposition requires the extra stresses of gravity Nevertheless, it appears that with an appropriate exercise program the physiological effects of weightlessness will not be a serious problem even during prolonged space voyages

REFERENCES

Adey, W R. The physiology of weightlessness. *Physiologist, 16*(1) 1974

American Physiological Society High Altitude and Man. D C, American Physiological Society, 1984

Armstrong, H G. Aerospace Medicine Baltimore, Williams & W 1961

Blomqvist, C G, and Stone, H L. Cardiovascular gravitational stress *In* Shepherd, J T., and Abboud, F M (eds Handbook of Physiology Sec 2, Vol III Bethesda, American Physiological Society, 1983, p 1025

Brendel, W (ed.) High Altitude Physiology and Medicine I Physiology of Adaptation New York, Springer Verlag, 1982

Bullard, R W Physiological problems of space travel *Annu Rev Physiol*, *34* 205, 1972

Burton, R R et al Man at high sustained +Gz acceleration. A review *Aerospace Med*, *45* 1115, 1974

Cardiovascular Problems Associated with Aviation Safety Springfield Va, National Technical Information Service, 1976

Ciba Foundation Symposia High Altitude Physiology New York Churchill Livingstone, 1971

Dill D B, *et al* Hemoconcentration at altitude *J Appl. Physiol.* *27* 514, 1969

Frisancho, A R Functional adaptation to high altitude hypoxia *Science 187* 313 1975

Grover, R F, *et al* High altitude pulmonary edema *In* Fishman, A P and Renkin E M (eds) Pulmonary Edema. Baltimore Waverly Press, 1979, p 229

Hempleman, H V, and Lockwood, A. P M. The Physiology of Diving in Man and Other Animals London, Edward Arnold, 1978

Hochachka, P W Living Without Oxygen Closed and Open Systems of Hypoxia Tolerance Cambridge, Mass, Harvard University Press, 1980

Hock, R J The physiology of high altitude *Sci. Am*, *222* 52, 1979

Kellogg, R H Altitude acclimatization, a historical introduction emphasizing the regulation of breathing *Physiologist*, *11* 37, 1968

Korner, P I Circulatory adaptations in hypoxia *Physiol. Rev*, *39* 687 1959

Lahiri S Physiological responses and adaptations to high altitude *In* MTP International Review of Science Physiology Vol 7. Baltimore University Park Press, 1974, p 271

Lahiri S Physiological responses and adaptations to high altitude *Ann Rev Physiol*, *15* 217, 1977

Lee, D H K (ed) Physiology, Environment, and Man. New York Academic Press, 1970

McCally, M Hypodynamics and Hypogravics. The Physiology of Inactivity and Weightlessness New York, Academic Press, 1969

Pace N Respiration at high altitude *Fed Proc*, *33* 2126, 1974

Private Pilot's Medical Handbook Blue Ridge Summit, Pa, TAB Books 1982

Read K Aeromedicine for Aviators New York, State Mutual Book and Periodical Services 1982

Reeves, J T, *et al* Physiological effects of high altitude on the pulmonary circulation *In* Robertshaw, D (ed) International Review of Physiology Environmental Physiology III Vol 20 Baltimore, University Park Press, 1979, p 289

Sandler H, and Winter, D L. Physiological Responses of Women to Simulated Weightlessness A Review of the Significant Findings of the First Female Bed Rest Study Springfield, Va, National Technical Information Service, 1978

Sloan, A W Man in Extreme Environments Springfield, Ill., Charles C Thomas, 1979

Smith, E E, and Crowell, J W Influence of hematocrit ratio on survival of unacclimatized dogs at simulated high altitude *Am. J Physiol*, *205* 1172, 1963

Smith, E E, and Crowell, J W Role of the hematocrit in altitude acclimatization *Aerospace Med*, *38* 39, 1966

Stickney, J. C Some problems of homeostasis in high altitude exposure *Physiologist*, *15* 349, 1972

Sutton, J R, et al., Exercise at altitude *Annu Rev Physiol*, *45* 427, 1983

45

Physiology of Deep Sea Diving and Other Hyperbaric Conditions

When human beings descend beneath the sea, the pressure around them increases tremendously To keep the lungs from collapsing, air must be supplied also under high pressure, which exposes the blood in the lungs to extremely high alveolar gas pressures and which is called *hyperbarism* Beyond certain limits these high pressures can cause tremendous alterations in the physiology of the body, which explains the necessity for the present discussion.

Also exposed to high atmospheric pressures are caisson workers who, in digging tunnels beneath rivers or elsewhere, often must work in a pressurized area to keep the tunnel from caving in. Here again, the same problems of excessively high gas pressures in the alveoli occur.

Before explaining the effects of hyperbarism on the body, it is necessary to review some physical principles of pressure and volume changes at different depths beneath the sea.

Relationship of Sea Depth to Pressure. A column of fresh water 34 feet high (sea water, 33 feet) exerts the same pressure at its bottom as all the atmosphere above the earth. Therefore, a person 33 feet beneath the ocean surface is exposed to a pressure of 2 atmospheres, 1 atmosphere of pressure caused by the air above the water and the second atmosphere by the weight of the water itself. At 66 feet the pressure is 3 atmospheres, and so forth, in accord with the table in Figure 45–1.

Effect of Depth on the Volume of Gases—Boyle's Law. Another important effect of depth is the compression of gases to smaller and smaller volumes Figure 45–1 also illustrates a bell jar at sea level containing 1 liter of air. At 33 feet beneath the sea where the pressure is 2 atmospheres, the volume has been compressed to only one-half liter. At 100 feet, where the pressure is 4 atmospheres, the volume has been compressed to one-fourth liter, and at 8 atmospheres (233 feet) to one-eighth liter. Thus, the volume to which a given quantity of gas is compressed is inversely proportional to the pressure. This is the physical principle called *Boyle's Law*, which is extremely important in diving because the increased pressures can cause the air chambers of the diver's body, including the lungs, to become so small in some instances that serious damage results, as is discussed later in the chapter.

Many times in this chapter it is necessary to refer to *actual volume* versus *sea level volume* For instance, we might speak of an actual volume of 1 liter at a depth of 300 feet, this is the same quantity of air as a sea level volume of 10 liters.

EFFECT OF HIGH PARTIAL PRESSURES OF GASES ON THE BODY

The three gases to which a diver breathing air is normally exposed are nitrogen, oxygen, and carbon dioxide. However, helium is often substituted for nitrogen in the diving mixture; therefore, the effects of this gas under high pressure must also be considered

Nitrogen Narcosis at High Nitrogen Pressures. Approximately four fifths of the air is nitrogen. At sea level pressure the nitrogen has no known effect on bodily function, but at high pressures it can cause varying degrees of narcosis. When the diver remains beneath the sea for an hour or more and is breathing compressed air, the depth at which the first symptoms of mild narcosis appear is approximately 120 feet, at which level he begins to exhibit joviality and to lose many of his cares. At 150 to 200 feet, he becomes drowsy. At 200 to 250 feet, his strength wanes considerably, and he often becomes too clumsy to perform the work required of him. Beyond 250 feet (8.5 atmospheres pressure), the diver usually becomes almost useless as a result of nitrogen narcosis if he remains at these depths too long.

Nitrogen narcosis has characteristics very similar to those of alcohol intoxication, and for this reason it has frequently been called "raptures of the depths."

The mechanism of the narcotic effect is believed to be the same as that of essentially all the gas anesthetics. That is, nitrogen dissolves freely in the fats of the body, and it is presumed that it, like most other anesthetic gases, dissolves in the membranes or other lipid structures of the neurons and, because of its *physical* effect on altering electrical conductance of the membrane, reduces their excitability.

Oxygen Toxicity at High Pressures. *Effect of Extremely High Po_2s on Blood Oxygen Transport.*

Depth (feet)	*Atmosphere(s)*
Sea level	1
33	2
66	3
100	4
133	5
166	6
200	7
300	10
400	13
500	16

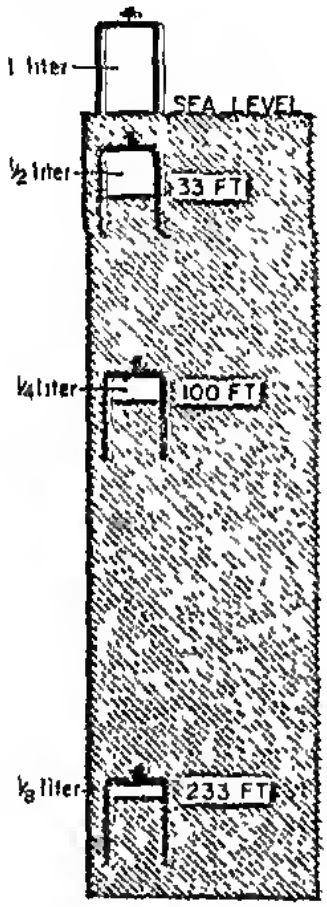


Figure 45–1. Effect of depth on gas volumes

When the Po_2 in the blood rises far above 100 mm Hg, the amount of oxygen dissolved in the water of the blood increases markedly. This is illustrated in Figure 45–2, which depicts the same oxygen-hemoglobin dissociation curve as that shown in Chapter 41 except that the alveolar Po_2 is now extended to over 3000 mm Hg. Also depicted is the volume of oxygen dissolved in the fluid of the blood at each Po_2 level. Note that in the normal range of alveolar Po_2, almost none of the total oxygen in the blood is accounted for by dissolved oxygen, but as the pressure rises progressively into the thousands of millimeters of mercury, a large portion of the total oxygen is then dissolved rather than bound with hemoglobin.

Effect of High Alveolar Po_2 on Tissue Po_2. Let us assume that the Po_2 in the lungs is about 3000 mm Hg (4 atmospheres pressure). Referring to Figure 45–2, one finds that this represents a total oxygen content in each 100 ml of blood of about 29 volumes per cent, as illustrated by Point A in the figure. As this blood pass through the tissue capillaries and the tissues utili their normal amount of oxygen, about 5 ml from ea 100 ml of blood, the oxygen content on leaving the tiss capillaries is still 24 volumes per cent (Point B in t figure). At this point the Po_2 is still about 1200 mm H which means that oxygen is delivered to the tissues this extremely high pressure instead of at the norm value of 40 mm Hg. Thus, once the alveolar Po_2 ris above a critical level, the hemoglobin-oxygen buff mechanism (discussed in Chapter 41) is no longer c pable of keeping the tissue Po_2 in the normal safe ran between 20 and 60 mm Hg.

Acute Oxygen Poisoning. Because of the extrem high tissue Po_2 that occurs when oxygen is breathed a very high alveolar oxygen pressure, one can read understand that this can be very detrimental to ma of the body's tissues. This is especially true of the bra In fact, exposure to 3 atmospheres pressure of oxyg (Po_2 = 2280 mm Hg) will cause convulsions followed coma in most persons after one to several hours. T convulsions often occur without any warning and, obvious reasons, are likely to be lethal to divers s merged beneath the sea.

Other symptoms encountered in acute oxygen poisi ing include *nausea, muscle twitchings, dizziness, disti bances of vision, irritability,* and *disorientation.*

Exercise greatly increases a diver's susceptibility oxygen toxicity, causing symptoms to appear much e lier and with far greater severity than in the rest person.

Excessive Intracellular Oxidation as the Cau of Oxygen Toxicity—"Oxidizing Free Radical Molecular oxygen (O_2) has very little capability of o dizing other chemical compounds. Instead, it must fi

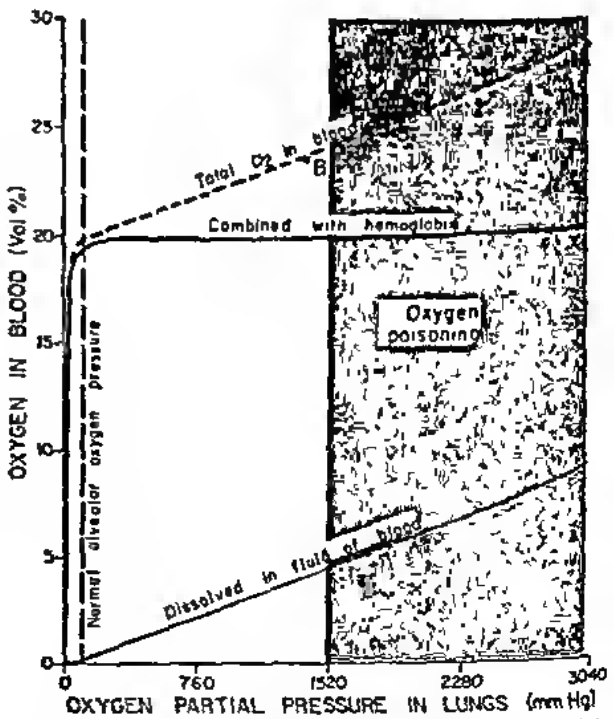


Figure 45–2. Quantity of oxygen dissolved in the fluid of the blood and in combination with hemoglobin at very high Po_2s.

be converted into an "active" form of oxygen. There are several different forms of active oxygen; these are often called *oxidizing free radicals*. One of the most important of these is the superoxide free radical O_2^-. Even when the tissue Po_2 is normal at the level of 40 mm Hg, small amounts of oxidizing free radicals are continually being formed from the dissolved molecular oxygen. However, the tissues also contain multiple enzymes that rapidly remove these free radicals, including especially peroxidases, catalases, and superoxide dismutases. Therefore, so long as the hemoglobin-oxygen buffering mechanism functions properly and maintains a normal tissue Po_2, the oxidizing free radicals are removed so rapidly that they have little or no effect in the tissues.

On the other hand, above a critical alveolar Po_2, the hemoglobin-oxygen buffering mechanism fails, and the tissue Po_2 can then rise to hundreds or thousands of millimeters of mercury. Then, the amounts of oxidizing free radicals literally swamp the enzyme systems for removing them, and now they do have serious destructive and even lethal effects on the cells. One of the principal effects is to oxidize the polyunsaturated fatty acids that are essential components of many of the membranous structures of the cells, and another effect is to oxidize some of the cellular enzymes, thus damaging severely the cellular metabolic systems. These effects can occur in all parts of the body, but the nervous tissues are especially susceptible because of their high lipid content. Therefore, most of the lethal effects of acute oxygen toxicity are related to brain dysfunction.

Another factor that may also play an important role in oxygen toxicity is the following: The very high tissue oxygen Po_2s cause severe arteriolar constriction, and the local tissue blood flow sometimes decreases to less than 50 per cent of normal. Under these conditions, it is possible that delivery of various cellular nutrients may become too low to maintain normal cellular function.

Chronic Oxygen Poisoning As a Cause of Pulmonary Disability. A person can be exposed to 100 per cent oxygen at normal atmospheric pressure almost *indefinitely* without developing the *acute* oxygen *toxicity* just described. However, after 12 or so hours of this exposure, *lung passageway congestion, pulmonary edema*, and *atelectasis* caused by damage to the linings of the bronchi and alveoli begin to develop. Here again, it appears that these local tissues of the lungs are damaged as a result of oxidative destruction of some of the essential elements of the tissues. The reason this effect occurs in the lungs and not in the other tissues at these lower alveolar Po_2s is that the lungs are directly exposed to the high oxygen pressure, whereas oxygen is delivered to the other tissues at essentially normal Po_2 because of the hemoglobin-oxygen buffer system described in Chapter 41. When the Po_2 in the air rises above about 1500 mm Hg, this buffer system fails, which then allows the Po_2 of all tissues to rise and to cause the acute oxygen poisoning described in the previous section.

Carbon Dioxide Toxicity at Great Depths. If the diving gear is properly designed and also functions properly, the diver has no problem from carbon dioxide toxicity, for depth alone does not increase the carbon dioxide partial pressure in the alveoli. This is true because depth does not increase the rate of CO_2 production in the body, and as long as the diver continues to breathe a normal tidal volume, he continues to expire the carbon dioxide as it is formed, maintaining his alveolar carbon dioxide partial pressure at a normal value.

Unfortunately, though, in certain types of diving gear, such as the diving helmet and the different types of rebreathing apparatuses, carbon dioxide can frequently build up in the dead space air of the apparatus and be rebreathed by the diver. Up to an alveolar carbon dioxide pressure (Pco_2) of about 80 mm Hg, two times that of normal alveoli, the diver tolerates this buildup, his minute respiratory volume increasing up to a maximum of eight- to elevenfold to compensate for the increased carbon dioxide. However, beyond the 80 mm Hg level the situation becomes intolerable, and eventually the respiratory center begins to be depressed rather than excited; the diver's respiration then begins to fail rather than to compensate. In addition, the diver develops severe respiratory acidosis, and varying degrees of lethargy, narcosis, and finally anesthesia ensue, as was discussed in Chapter 42.

Effects of Helium at High Pressures. In very deep dives, helium is used in place of nitrogen because it has only one fourth to one fifth the narcotic effect of nitrogen. Almost no effects of helium are observed until a person descends below about 500 feet. Beyond this level divers begin to experience a so-called "high pressure nervous syndrome" which is characterized by tremor at about 700 feet, drowsiness from 1000 feet onward, and lack of coordination at still deeper levels.

Aside from the much smaller narcotic effect of helium than nitrogen on the nervous system, three other properties also make it desirable in the diving gas mixture under some conditions:

(1) Because of its low atomic weight, its density is slight, which reduces the airway resistance of the diver. (2) Also because of its low atomic weight, helium diffuses through the tissues much more rapidly than nitrogen, which allows more rapid removal of helium than nitrogen from the body fluids. (3) Helium is less soluble in the body fluids than nitrogen, which reduces the quantity of bubbles that can form in the tissues when the diver is decompressed after a prolonged dive.

DECOMPRESSION OF THE DIVER AFTER EXPOSURE TO HIGH PRESSURES

When a person breathes air under high pressure for a long time, the amount of nitrogen dissolved in the body fluids becomes great. The reason for this is the following: The blood flowing through the pulmonary capillaries becomes saturated with nitrogen to the same pressure as that in the breathing mixture. Over several hours, enough nitrogen is carried to all the tissues of the body to saturate the tissues also with dissolved nitrogen. And, since nitrogen is not metabolized by the body, it remains dissolved until the nitrogen pressure in the lungs decreases, at which time the nitrogen is then removed by the reverse respiratory process.

Volume of Nitrogen Dissolved in the Body Fluids at Different Depths. At sea level almost 1 liter of nitrogen is dissolved in the entire body. A little less than half of this is dissolved in the water of the body and a little more than half in the fat of the body. This is true despite the fact that fat constitutes only 15 per cent of the normal body because nitrogen is five times as soluble in fat as in water.

After the diver has become totally saturated with

nitrogen the *sea level volume of nitrogen* dissolved in the body fluids at the different depths is:

feet	*liters*
33	2
100	4
200	7
300	10

However, several hours are required for the gas pressures of nitrogen in all the body tissues to come to equilibrium with the gas pressure of nitrogen in the alveoli simply because the blood does not flow rapidly enough and the nitrogen does not diffuse rapidly enough to cause instantaneous equilibrium. The nitrogen dissolved in the water of the body comes to almost complete saturation in about one hour, but the fat, requiring much more nitrogen for saturation and also having a relatively poor blood supply, reaches saturation only after several hours. For this reason, if a person remains at deep levels for only a few minutes not much nitrogen dissolves in fluids and tissues, whereas if the person remains at a deep level for several hours fluids and tissues become almost completely saturated with nitrogen.

Decompression Sickness (Synonyms: Bends, Compressed Air Sickness, Caisson Disease, Diver's Paralysis, Dysbarism). If a diver has been beneath the sea long enough so that large amounts of nitrogen have dissolved in his body and then he suddenly comes back to the surface of the sea, significant quantities of nitrogen bubbles can develop in his body fluids either intracellularly or extracellularly, and these can cause minor or serious damage in almost any area of the body, depending on the amount of bubbles formed; this is decompression sickness.

The principles underlying bubble formation are shown in Figure 45–3. To the left, the diver's tissues have become equilibrated to a very high nitrogen pressure. However, as long as the diver remains deep beneath the sea, the pressure against the outside of his body (5000 mm Hg) compresses all the body tissues sufficiently to keep the dissolved gases in solution. But when the diver suddenly rises to sea level, the pressure on the outside of his body becomes only 1 atmosphere (760 mm Hg), whereas the pressure inside the body fluids is the sum of the pressures of water vapor, carbon dioxide, oxygen and nitrogen, or a total of 4065 mm Hg, which is greater than the pressure on the outside of the body. Therefore, the gases can escape from the dissolved state and form actual bubbles inside the tissues. However, the bubbles may not appear for many minutes to hours because the gases can remain dissolved in the "supersaturated" state sometimes for hours before bubbling.

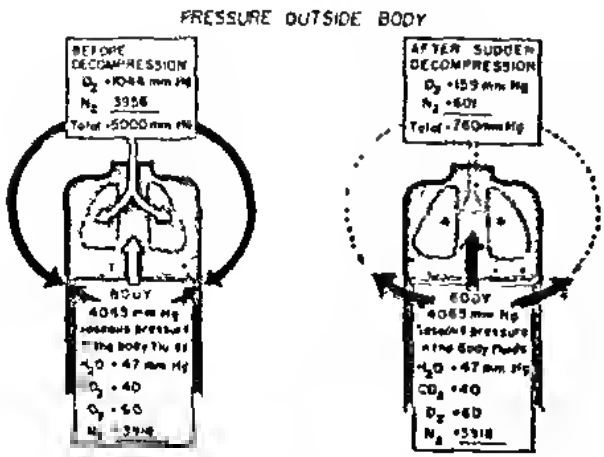


Figure 45–3. Gaseous pressures both inside and outside the body showing at right the great excess of intrabody pressure that is responsible for bubble formation in the body tissues.

Exercise hastens the formation of bubbles during decompression because of increased agitation of the tissues and fluids. This is an effect analogous to that of shaking an opened bottle of soda pop to release the bubbles.

Symptoms of Decompression Sickness. In persons who have developed decompression sickness, symptoms have occurred with the following frequencies:

	per cent
Local pain in the legs or arms	89
Dizziness	5.3
Paralysis	2.3
Shortness of breath ("the chokes")	1.6
Extreme fatigue and pain	1.3
Collapse with unconsciousness	0.5

From this list it can be seen that the most serious problems are usually related to bubble formation in the nervous system. Bubbles sometimes actually disrupt important pathways in the brain or spinal cord, the bubbles in the peripheral nerves can cause severe pain. Unfortunately, formation of large bubbles in the central nervous system occasionally even leads to permanent paralysis or permanent mental disturbances.

But the nervous system is not the only locus of damage in decompression sickness, for bubbles can also form in the blood and become caught in the capillaries of the lungs; these bubbles block pulmonary blood flow and cause "the chokes," characterized by serious shortness of breath. This is often followed by severe pulmonary edema, which further aggravates the condition and can cause death.

The symptoms of decompression sickness usually appear within a few minutes to an hour after sudden decompression. However, occasional symptoms of decompression sickness develop as long as six or more hours after decompression.

Rate of Nitrogen Elimination from the Body, Decompression Tables. Fortunately, if a diver is brought to the surface slowly, the dissolved nitrogen is eliminated through the lungs rapidly enough to prevent decompression sickness. Approximately two thirds of the total nitrogen is liberated in one hour and about 90 per cent in six hours. However, some excess nitrogen is still present for many hours, especially in the fatty tissues such as the brain, and the diver is not completely safe for as long as 9 to 12 hours. Therefore, a diver must be decompressed over a period of many hours if he has been deep beneath the sea for a long time. This decompression is achieved by bringing the diver to the surface slowly, over a period of many minutes to hours.

The rate at which a diver can be brought to the surface depends on, first the *depth* to which he has descended and, second, the *amount of time* he has been there. If he remains at deep levels for only a short time the body fluids do not become saturated, and, therefore the decompression time can be reduced accordingly.

Special decompression tables have been prepared by the U.S. Navy that detail procedures for safe decompression. To give the student an idea of the decompression

process, a diver who has been breathing air and has been on the bottom for 60 minutes at a depth of 190 feet is decompressed according to the following schedule:

10 minutes at 50 feet depth
17 minutes at 40 feet depth
19 minutes at 30 feet depth
50 minutes at 20 feet depth
84 minutes at 10 feet depth

Thus, for a work period on the bottom of only one hour, the total time for decompression is about three hours.

Decompression in a Tank and Treatment of Decompression Sickness. Another procedure for decompression, used especially in heavily polluted waters and when climatic situations require it, involves bringing the diver to the surface immediately and then placing him in a decompression tank within five minutes after arriving at the surface. Pressure is reapplied, and an appropriate decompression table that prevents bubble formation is used.

A person who begins to develop symptoms of decompression sickness can also be treated by being placed in such a decompression tank for a long time, several times as long as the usual decompression times, and allowing the nitrogen to be released from the body slowly. Also, for best results, the percentage of oxygen in the gas mixture of the decompression tank is increased, sometimes to pure oxygen. This reduces the nitrogen partial pressures in the alveoli to very low values, which causes several times as rapid removal of the nitrogen.

Use of Helium-Oxygen Mixtures in Deep Dives. In deep dives, helium has advantages over nitrogen, including (1) decreased decompression time, (2) lack of narcotic effect, and (3) decreased airway resistance in the lungs. The decreased decompression time results from two of its properties: (1) Only 40 per cent as much helium dissolves in the body as does nitrogen (2) Because of its small atomic size it diffuses through the tissues at a velocity about 2 5 times that of nitrogen.

However, helium is not always more advantageous than nitrogen in the breathing mixture For instance, for short and shallow dives, the rapid rate of diffusion of helium allows for more rapid buildup of dissolved helium in the tissue fluids than is true for nitrogen, which has a much slower rate of diffusion Therefore, the amount of dissolved gas is actually less when using nitrogen than when using helium. In general, if the bottom time is one hour, nitrogen is more advantageous at depths less than 100 feet and helium more advantageous at greater depths

SOME PHYSICAL PROBLEMS OF DIVING

Aside from the effects of high gas pressures on the body, still other physical factors place limitations on diving. *For instance*

"Sea Level" Volume of Air That the Diver Must Breathe to Eliminate Carbon Dioxide. To maintain the P_{CO_2} in the alveoli at a normal level, the *tidal volume* of air flowing in and out of the lungs with each breath must remain the same regardless of the depth of the dive. A tidal volume of 0.5 liter at 300 feet deep (10 atmospheres pressure) contains the same *quantity* of air as a sea level volume of 5 liters. Therefore, when the diver is working at a depth of 300 feet, the total *quantity* of air, *expressed in sea level volumes*, must be ten times as great as would be required when the same diver is working at the surface of the sea. To express this still another way, the diver exhausts the quantity in his tank at a rate ten times normal when working 300 feet below the sea, and this seriously reduces the amount of time that his breathing mixture will last. Thus, the typical sea level volumes of air that a worker requires each minute at different depths to blow off the required amounts of carbon dioxide are approximately the following

feet	*cubic feet*
Sea level	1 5
33	3
66	4.5
100	6
200	10 5
300	15

Change in Density of the Air—Effect on Maximum Breathing Capacity. The density of the air increases in proportion to the pressure, which means that the density is four times as great at 100 feet depth as at sea level and seven times as great at 200 feet.

The resistance to air flow through the respiratory passageways increases directly in proportion to the density of the breathing mixture Therefore, one can readily see that the increased density of the air will increase the work of breathing and, as a corollary, will decrease the maximum breathing capacity (the amount of air that one can breathe each minute) The following table gives the *maximum breathing capacity* in per cent of normal at different depths when breathing air and when breathing a mixture in which helium replaces the nitrogen of the air

Depth (ft)	Air (% of normal)	Oxygen-helium (% of normal)
25	75	100+
50	60	100+
100	50	86
200	35	63
400	24	48

Effect of Rapid Descent—"The Squeeze." On rapid descent, the volumes of all gases in the body become greatly reduced because of increasing pressure applied to the outside of the body. If additional quantities of air are supplied to the gas cavities during descent—especially to the lungs—no harm is done, but, if the person continues to descend without addition of gas to the cavities, the volume becomes greatly reduced, and serious physical damage results, this is called "the squeeze."

The most damaging effects of the squeeze occur in the lungs, for the smallest volume that the lungs can normally achieve is approximately 1.2 liters. Even if the *diver inspires a maximal breath prior to descending*, he can go down no farther than 100 feet before his chest begins to cave in. Therefore, to prevent lung squeeze, the diver must inspire additional air as he descends.

When air becomes entrapped in the middle ear during descent, the squeeze can cause a ruptured tympanum or bleeding into the middle ear, and when air is entrapped in one of the nasal sinuses intense pain and sinus

bleeding results. Occasionally, also, when a diver loses air pressure in his mask, the eyes can actually pop out into the mask and the face can become greatly distorted to fill the mask.

Overexpansion of the Lungs on Rapid Ascent—Air Embolism. Exactly the opposite pulmonary effects occur on rapid ascent if the person fails to expel air from the lungs on the way up Unfortunately, panic can frequently cause a person to close his glottis spastically, which can result in serious damage to the lungs. When the lungs become expanded to their limit the pressure in the alveoli then becomes much greater than the external pressure, and above an *excess* alveolar pressure of 80 to 100 mm Hg, air is forced into the pulmonary capillaries, causing air embolism in the circulation and often resulting in death. Also, increased pressure in the lungs frequently blows out large blebs on the surfaces of the lungs or ruptures the lungs to cause pneumothorax

In rare instances, in a diver who has been deep below the sea for a long time and in whom large amounts of gas have accumulated in the abdomen, rapid ascent can also cause serious trauma in the gastrointestinal tract

Rapid ascent can be especially serious when a person is attempting to escape from a submarine and must rise to the surface rapidly without appropriate diving gear It also occurs frequently when the diver loses control of his diving gear and his suit balloons up so greatly that he "blows up" to the surface

SCUBA DIVING (SELF-CONTAINED UNDERWATER BREATHING APPARATUS)

Prior to the 1940s, almost all diving was done using a diving helmet connected to a hose through which air was pumped to the diver from the surface Then, in 1943, Jacques Cousteau developed and popularized the *self-contained underwater breathing apparatus* which is popularly known simply as the SCUBA apparatus There are two basic types of SCUBA apparatuses (1) the *open circuit demand system*, which is used in perhaps 99 per cent of all sports and commercial diving; and (2) the *closed circuit system*, which is mainly experimental

The Open Circuit Demand System. Figure 45–4 illustrates an open circuit demand type of underwater breathing apparatus showing the following components (1) one or more tanks of compressed air or of some other breathing mixture, (2) a first stage "reducing" valve for reducing the pressure from the tanks to a constant low pressure level, (3) a combination inhalation "demand" valve and exhalation valve that allows air to be pulled into the lungs with very slight negative pressure and then to be exhausted into the sea, and (4) a mask and tube system with small "dead space."

Basically, the demand system operates as follows The first stage reducing valve reduces the pressure from the tanks usually to a pressure of about 140 lbs per square inch However, the breathing mixture does not flow continually into the mask Instead, with each inspiration, slight negative pressure in the mask pulls the diaphragm of the demand valve inward, and this automatically releases air from the hose into the mask and lungs In this way only the amount of air needed for inhalation enters the system Then, on expiration, the

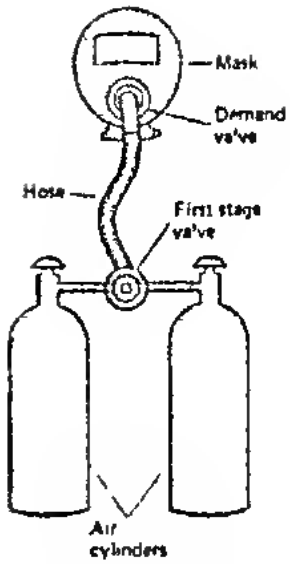


Figure 45–4. The open circuit demand type of SCUBA apparatus.

air cannot go back into the tank but instead is expired through the expiration valve.

The most important problem in use of the self-contained underwater breathing apparatus is the time limit that one can remain beneath the surface; only a few minutes are possible at great depths because tremendous airflow from the tanks is required to wash carbon dioxide out of the lungs—the greater the depth, the greater the airflow in terms of *quantity* of air that is required, as discussed earlier

The Closed Circuit System. In the simplest type of closed circuit system a person breathes pure oxygen. This system contains the following elements (1) a tank of pure oxygen, (2) a rubber bellows into which the diver can breathe back and forth, (3) a valve system for allowing oxygen to flow from the oxygen tank into the bellows as needed to keep it moderately filled, (4) a canister containing soda lime through which the rebreathed air passes to absorb the carbon dioxide, and (5) an appropriate mask system with valves to keep the gaseous mixture flowing in the right direction through the canister for carbon dioxide removal. Thus, the closed circuit system is similar to a standard anesthetic machine in which oxygen is continually rebreathed—except that no anesthetic is used.

The most important problem in use of this oxygen-filled closed circuit system is limitation in depth at which a person can remain beneath the sea because of poor oxygen tolerance by the body A safe working rule is no more than 30 minutes at 30 feet depth.

Several new closed circuit systems are available on an experimental basis. These utilize mixtures of oxygen and helium, and they have separate supply tanks for each gas. As oxygen is utilized from the breathing mixture, an electrical oxygen sensor and control mechanism replenishes the oxygen from the oxygen tank but keeps the oxygen concentration from rising too high; at the same time carbon dioxide is removed by a carbon dioxide absorbent. Since this gaseous mixture can be rebreathed again and again without ever blowing off any of it into the sea, none of the gases are wasted. Therefore, very small tanks of oxygen and helium allow the diver to remain under the sea for many hours.

SPECIAL PHYSIOLOGICAL PROBLEMS IN SUBMARINES

Escape from Submarines. Essentially the same problems as those of deep sea diving are often met in relation to submarines, especially when it is necessary to escape from a submerged submarine. Escape is possible from as deep as 300 feet even without using any special type of apparatus. Proper use of rebreathing devices using helium or hydrogen theoretically can allow escape from as deep as 600 feet or perhaps more.

One of the major problems of escape is prevention of air embolism. As the person ascends, the gases in his lungs expand and sometimes rupture a pulmonary vessel, allowing the gases to enter into the pulmonary vascular system to cause embolism of the circulation. Therefore, as the person ascends, he must exhale continually.

Expansion and exhalation of gases from the lungs during ascent, even without breathing, is often rapid enough to blow off the accumulating carbon dioxide in the lungs. This keeps the concentration of carbon dioxide from building up in the blood and keeps the person from having the desire to breathe. Therefore, he can hold his breath for an extra long time during ascent.

Health Problems in the Submarine Internal Environment. Except for escape, submarine medicine generally centers around several engineering problems to keep hazards out of the internal environment of the submarine. In atomic submarines there exists the problem of radiation hazards, but, with appropriate shielding, the amount of radiation received by the crew submerged beneath the sea has actually been less than the normal radiation received above the surface of the sea from cosmic rays. Therefore, no essential hazard results from this unless some failure of the apparatus causes unexpected release of radioactive materials.

Second, poisonous gases on occasion escape into the atmosphere of the submarine and must be controlled early. For instance, during several weeks' submergence, cigarette smoking by the crew can liberate sufficient *amounts of carbon monoxide, if it is not removed from* the air, to cause carbon monoxide poisoning, and on occasion even Freon gas has been found to diffuse through the tubes in refrigeration systems in sufficient quantity to cause toxicity.

HYPERBARIC OXYGEN THERAPY

In recent years it has been learned that the intense oxidizing properties of high pressure oxygen *(hyperbaric oxygen)* can have very valuable therapeutic effects in several important clinical conditions. Therefore, large pressure tanks are now available in many medical centers into which patients can be placed and treated with hyperbaric oxygen. The oxygen is usually administered at P_{O_2}s of 2 to 3 atmospheres of pressure. It is believed that the same oxidizing free radicals responsible for oxygen toxicity are also responsible for the therapeutic benefits. Some of the conditions in which hyperbaric oxygen therapy has been especially beneficial are the following:

Probably the most successful use of hyperbaric oxygen has been in the treatment of *gas gangrene*. The bacteria that cause this condition, the *clostridial organisms*, grow best under anaerobic conditions and actually stop growing at oxygen pressures greater than about 70 mm Hg. Therefore, hyperbaric oxygenation of the tissues can frequently stop the infectious process entirely and thus convert a condition that formerly was amost 100 per cent fatal into one that is cured in most instances by hyperbaric therapy.

Recent results also suggest that hyperbaric oxygenation might have almost as dramatic an effect in curing *leprosy* as in curing gas gangrene—also because of the susceptibility of the leprosy bacillus to destruction by high oxygen pressures.

Other conditions in which hyperbaric oxygen therapy has been either valuable or possibly valuable include decompression sickness, arterial gas embolism, carbon monoxide poisoning, osteomyelitis, and possibly myocardial infarction.

REFERENCES

Anderson, H T Physiological adaptations in diving vertebrates *Physiol Rev*, *46* 212, 1966

Behnke A R, Jr, and Lanphier, E H Underwater physiology. *In* Fenn, W O, and Rahn, H (eds) Handbook of Physiology Sec 3, Vol 2, Baltimore, Williams & Wilkins, 1965, p 1159

Bennett, P B, and Elliott, D H The Physiology and Medicine of Diving and Compressed Air Work, 2nd Ed Baltimore, Williams & Wilkins, 1975

Burakovsky, V I Hyperbaric Oxygenation and Its Value in Cardiovascular Surgery Chicago, Imported Publications, 1981

Fisher, A B, *et al* Oxygen toxicity of the lung Biochemical aspects *In* Fishman, A P, and Renkin, E M (eds) Pulmonary Edema Baltimore, Waverly Press, 1979, p 207

Fridovich, I Superoxide radical An endogenous toxicant *Annu Rev. Pharmacol Toxicol*, *23*:239, 1983

Gamarra, J A Decompression Sickness Hagerstown, Md, Harper & Row 1974

Gooden, B A Drowning and the diving reflex in man *Med. J. Aust*, *2* 583, 1972.

Halsey, M J Effects of high pressure on the central nervous system *Physiol Rev*, *62* 1341, 1982.

Hochachka, B W, and Murphy, B, Metabolic status during diving and recovery in marine mammals *In* Robertshaw, D (ed) International Review of Physiology Environmental Physiology III Vol 20, Baltimore, University Park Press, 1979, p 253

Hochachka, B W, and Storey, K B Metabolic consequences of diving in animals and man *Science*, *187* 613, 1975

Miles, S, Underwater Medicine, 2nd Ed Philadelphia, J. B Lippincott Co, 1966

Miller, J. W Vertical Excursions Breathing Air from Nitrogen Oxygen or Air Saturation Exposures Washington, D C, U S Government Printing Office, 1976

Oxygen Free Radicals and Tissue Damage Ciba Foundation Symposium New York, Excerpta Medica, 1979

Shilling, C W, and Beckett, M W (eds) Underwater Physiology IV, Bethesda, Md, Federation of American Societies for Experimental Biology, 1978.

Sloan, A. W Man in Extreme Environments, Springfield, Ill, Charles C Thomas, 1979

Undersea Medical Society Glossary of Diving and Hyperbaric Terms. Bethesda, Md, The Society, 1978

Vail, E G Hyperbaric respiratory mechanics *Aerospace Med.*, *42*:536, 1971.

Zimmerman, A M. (ed): High Pressure Effects on Cellular Processes. New York, Academic Press, 1970

IX

THE NERVOUS SYSTEM

46 ■ Organization of the Nervous System; Basic Functions of Synapses

47 ■ Neuronal Mechanisms and Circuits for Processing Information

48 ■ Sensory Receptors and Their Basic Mechanisms of Action

49 ■ Somatic Sensations: I. The Mechanoreceptive Sensations

50 ■ Somatic Sensations: II. Pain, Visceral Pain, Headache, and Thermal Sensations

51 ■ Motor Functions of the Spinal Cord; and the Cord Reflexes

52 ■ Motor Functions of the Brain Stem and Basal Ganglia—Reticular Formation, Vestibular Apparatus, Equilibrium, and Brain Stem Reflexes

53 ■ Cortical and Cerebellar Control of Motor Functions

54 ■ The Cerebral Cortex and Intellectual Functions of the Brain

55 ■ Activation of the Brain—The Reticular Activating System and Wakefulness; Attention and Searching of Memory; Brain Waves; Sleep; Epilepsy

56 ■ Behavioral Functions of the Brain: The Limbic System, Role of the Hypothalamus, and Control of Vegetative Functions of the Body

57 ■ The Autonomic Nervous System; The Adrenal Medulla

46

Organization of the Nervous System; Basic Functions of Synapses

The nervous system, along with the endocrine system, provides most of the control functions for the body. In general, the nervous system controls the rapid activities of the body, such as muscular contractions, rapidly changing visceral events, and even the rates of secretion of some endocrine glands The endocrine system, by contrast, regulates principally the metabolic functions of the body

Tha nervous system is unique in the vast complexity of the control actions that it can perform It receives literally millions of bits of information from the different sensory organs and then integrates all these to determine the response to be made by the body The purpose of this chapter is to present, first, a general outline of the overall mechanisms by which the nervous system performs such functions Then we will discuss the function of central nervous system synapses, the basic structures that control the passage of signals into, through, and then out of the nervous system. In succeeding chapters we will analyze in detail the functions of the individual parts of the nervous system Before beginning this discussion, however, the reader should refer to Chapters 10 and 12, which present, respectively, the principles of membrane potentials and transmission of signals through neuromuscular junctions.

GENERAL DESIGN OF THE NERVOUS SYSTEM

THE SENSORY DIVISION—SENSORY RECEPTORS

Most activities of the nervous system are initiated by sensory experience emanating from *sensory receptors,* whether these be visual receptors, auditory receptors, tactile receptors on the surface of the body, or other kinds of receptors. This sensory experience can cause an immediate reaction, or its memory can be stored in the brain for minutes, weeks, or years and then can help determine the bodily reactions at some future date.

Figure 46–1 illustrates a portion of the sensory system, the *somatic* portion that transmits sensory

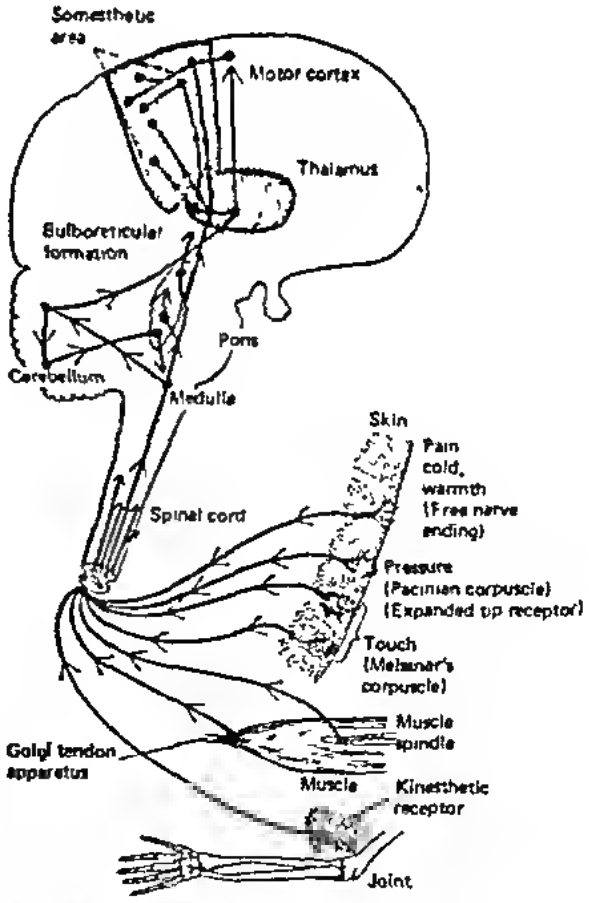


Figure 46–1. The somatic sensory axis of the nervous system.

information from the receptors of the entire surface of the body and deep structures. This information enters the central nervous system through the spinal nerves and is conducted to multiple "primary" sensory areas in (a) the spinal cord at all levels, (b) the reticular substance of the medulla, pons, and mesencephalon, (c) the cerebellum, (d) the thalamus, and (e) the somesthetic areas of the cerebral cortex. But in addition to these primary sensory oreas, signals are then relayed to essentially all other parts of the nervous system as well.

THE MOTOR DIVISION—THE EFFECTORS

The most important ultimate role of the nervous system is control of bodily activities. This is achieved by controlling (a) contraction of skeletal muscles throughout the body, (b) contraction of smooth muscle in the internal organs, and (c) secretion by both exocrine ond endocrine glands in many parts of tha body. These activities are collectively called *motor functions* of the nervous system, and the muscles and glands are called *effectors* because they perform the functions dictated by the nerve signala.

Figure 46–2 illustrates the *motor axis* of the nervous syetem for controlling skeletal muscle contraction. Operating parallel to this axie is another similar system for control of the smooth muscles and glands; it is the *autonomic nervous system*, which will be presented in Chapter 57.

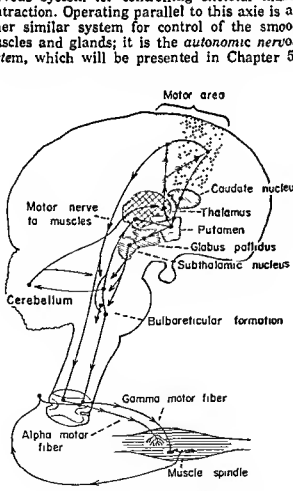


Figure 46–2. The motor axis of the nervous system.

Note in Figure 46–2 that the skeletal muscles can be controlled from many different levels of the central nervous system, including (a) tbe spinal cord, (b) tbe reticular substance of the medulla, pons, and mesencephalon, (c) the basal ganglia, (d) the cerebellum, and (e) the motor cortex. Each of these different areas plays its own specific role in the control of body movements, the lower regions being concerned primarily with automatic, instantaneous responses of tbe body to sensory stimuli and the higher regions with deliberate movements controlled by the thought processes of the cerebrum.

PROCESSING OF INFORMATION—"INTEGRATIVE" FUNCTION OF THE NERVOUS SYSTEM

The nervous system would not be at all effective in controlling bodily functions if each bit of sensory information caused some motor reaction. Therefore, one of the major functions of the nervous system is to process incoming information in such a way that *appropriate* motor responses occur. Indeed, more than 99 per cent of all senaory information is discarded by the brain as irrelavant and unimportant. For instance, one ia ordinarily totally unaware of the parts of the body that are in contact with clothing and is also unawara of the seat pressure when eitting. Likewise, attention is drawn only to an occasional object in one'e field of vision, and even the perpetual noise of our surroundings is usually relegated to the background.

After the important sensory information haa been selected, it is then channeled into proper motor regions of the brain to causa the desired responses, which is called the *integrative function* of the nervous system. Thus, if a person places a hand on a hot stove, the desired response is to lift the hand, plus other associated responses such as moving the entire body away from the stove and perhaps even shouting with pain. Yet even these responses represent activity by only a small fraction of the total motor system of the body.

Role of Synapses in Processing Information. The synapse is the junction point from one neuron to the next and, therefore, is an advantageous site for control of signal transmission. Later in this chapter we will discuss the details of synaptic function. However, it is important to point out here that the synapses determine the directions that the nervous signals spread in the nervous system. Some synapses transmit signals from one neuron to the next with ease, while others transmit signals only with difficulty. Also, facilitatory and inhibitory signals from other areas in the nervous system can control synaptic activity, sometimes opening the synapses for transmission and other times closing them. In addition, some postsynaptic neurons respond with large numbers of impulses, while others respond with only a few. Thus, the

synapses perform a selective action, often blocking the weak signals while allowing the strong signals to pass, often selecting and amplifying certsin weak signals, and often channeling the signal in many different directions rather than simply in one direction

STORAGE OF INFORMATION—MEMORY

Only a small fraction of the important sensory information causes an immediate motor response. Much of the remainder is stored for future control of motor activities and for use in the thinking processes Most of this storage occurs in the *cerebral cortex*, but not all, for even the basal regions of the brain and perhaps even the spinal cord can store small amounts of information.

The storage of information is the process we call *memory*, and this too is a function of the synapses That is, each time certain types of sensory signals, pass through sequences of synapses, these synapses become more capable of transmitting the same signals the next time, which process is called *facilitation*. After the sensory signals have passed through the synapses a large number of times, the synapses become so facilitated that signals generated within the brain itself can also cause transmission of impulses through the same sequences of synapses even though the sensory input has not been excited This gives the person a perception of experiencing the original sensations, though in effect they are only memories of the sensations.

Unfortunately, we do not know the precise mechanism by which facilitation of synapses occurs in the memory process, but what is known about this and other details of the memory process will be discussed in Chapter 54

Once memories have been stored in the nervous system, they become part of the processing mechanism. The thought processes of the brain compare new sensory experiences with the stored memories; the memories help to select the important new sensory information and to channel this into appropriate storage areas for future use or into motor areas to cause bodily responses

THE THREE MAJOR LEVELS OF CENTRAL NERVOUS SYSTEM FUNCTION

The human nervous system has inherited specific characteristics from each stage of evolutionary development From this heritage, three major levels of the central nervous system have specific functional attributes: (1) the *spinal cord level*, (2) the *lower brain level*, and (3) the *higher brain* or *cortical level*

The Spinal Cord Level

We often think of the spinal cord as being only [illegible]ut for signals from the periphery of the body to the brain or in the opposite direction from the brain back to the body. However, this is far from the truth. Even after the spinal cord has been cut in the high neck region, many spinal cord functions still occur. For instance, neuronal circuits in the cord can cause

(1) walking movements,

(2) reflexes that withdraw portions of the body from objects,

(3) reflexes that stiffen the legs to support the body against gravity,

(4) reflexes that control local blood vessels, gastrointestinal movements, and so forth,

(5) and many other functions.

In fact, the upper levels of the nervous system often operate not by sending signals directly to the periphery of the body but instead by sending signals to the control centers of the cord, simply "commanding" the cord centers to perform their functions

The Lower Brain Level

Many, if not most of what we call subconscious activities of the body are controlled in the lower areas of the brain—in the medulla, pons, mesencephalon, hypothalamus, thalamus, cerebellum, and basal ganglia Subconscious control of arterial pressure and respiration is achieved mainly in the medulla and pons Control of equilibrium is a combined function of the older portions of the cerebellum and the reticular substance of the medulla, pons, and mesencephalon Feeding reflexes, such as salivation in response to the taste of food and the licking of the lips, are controlled by areas in the medulla, pons, mesencephalon, amygdala, and hypothalamus; and many emotional patterns, such as anger, excitement, sexual activities, reaction to pain, or reaction of pleasure, can occur in animals without a cerebral cortex.

The Higher Brain or Cortical Level

After recounting all the nervous system functions that can occur at the cord and lower brain levels, what is left for the cerebral cortex to do? The answer to this is a complex one, but it begins with the fact that the cerebral cortex is an extremely large memory storehouse. The cortex never functions alone but always in association with the lower centers of the nervous system.

Without the cerebral cortex, the functions of the lower brain centers are often very imprecise The vast storehouse of cortical information usually converts these functions to very determinative and precise operations.

Finally, the cerebral cortex is essential for most of our thought processes even though it also cannot function alone in this. In fact, it is the lower centers that cause *wakefulness* in the cerebral cortex, thus opening its bank of memories to the thinking machinery of the brain.

Thus, each portion of the nervous system performs specific functions. Many integrative functions are well developed in the spinal cord, and many of the subconscious functions of the brain are originated and executed entirely in the lower regions of the brain. But it is the cortex that opens the world up for one's mind.

COMPARISON OF THE NERVOUS SYSTEM WITH AN ELECTRONIC COMPUTER

When electronic computers were first developed in many different laboratories of the world by as many different scientists, it soon became apparent that all these machines have many features in common with the nervous system. First, they all have input circuits that are comparable to the sensory portion of the nervous system and output circuits that are comparable to the motor portion of the nervous system. In the conducting pathway between the inputs and the outputs are the mechanisms for performing the different types of computations.

In simple computers, the output signals are controlled directly by the input signals, operating in a manner similar to that of the simple reflexes of the spinal cord. But, in the more complex computers, the output is determined both by the input signals and by information that has already been stored in memory in the computer, which is analogous to the more complex reflex and processing mechanisms of our higher nervous system. Furthermore, as the computers become even more complex it is necessary to add still another unit, called the *central programming unit*, which determines the sequence of all operations. This unit is analogous to the mechanism in our brain that allows us to direct our attention first to one thought or sensation or motor activity, then to another, and so forth, until complex sequences of thought or action take place.

Figure 46–3 illustrates a simple block diagram of a modern computer. Even a rapid study of this diagram will demonstrate its similarity to the nervous system. The fact that the basic components of the general purpose computer are analogous to those of the human nervous system demonstrates that the brain is basically a computer that continuously collects sensory information and uses this along with stored information to compute the daily course of bodily activity.

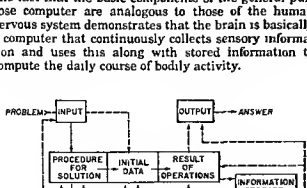


Figure 46–3. Block diagram of a general purpose electronic computer, showing the basic components and their interrelationships.

THE CENTRAL NERVOUS SYSTEM SYNAPSES

Every medical student is aware that information is transmitted in the central nervous system mainly in the form of nerve impulses through a succession of neurons, one after another. However, it is not immediately apparent that each impulse (a) may be blocked in its transmission from one neuron to the next, (b) may be changed from a single impulse into repetitive impulses, or (c) may be integrated with impulses from other neurons to cause highly intricate patterns of impulses in successive neurons. All these functions can be classified as *synaptic functions of neurons*.

Types of Synapses—Chemical and Electrical

Nerve signals are transmitted from one neuron to the next through interneuronal junctions called *synapses*. In the animal world there are basically two different types of synapses: (1) the *chemical synapse*, and (2) the *electrical synapse*.

From a practical view, almost all the synapses utilized for signal transmission in the central nervous system are chemical synapses. In these, the first neuron secretes a chemical substance called a *neurotransmitter* at the synapse, and this transmitter in turn acts on receptor proteins in the membrane of the next neuron to excite the neuron, to inhibit it, or to modify its sensitivity in some other way. Over 30 different transmitter substances have been discovered thus far.

Electrical synapses are characterized by direct channels that conduct electricity from one cell to the next. Most of these consist of small protein tubular structures called *gap junctions* that allow free movement of ions from the interior of one cell to the next. These were discussed in Chapter 9. Only a few gap junctions have been found in the central nervous system, and their significance is not known. On the other hand, it is by way of gap junctions and other similar junctions that action potentials are transmitted from one smooth muscle fiber to the next in visceral smooth muscle (Chapter 11) and also from one cardiac muscle cell to the next in cardiac muscle (Chapter 13).

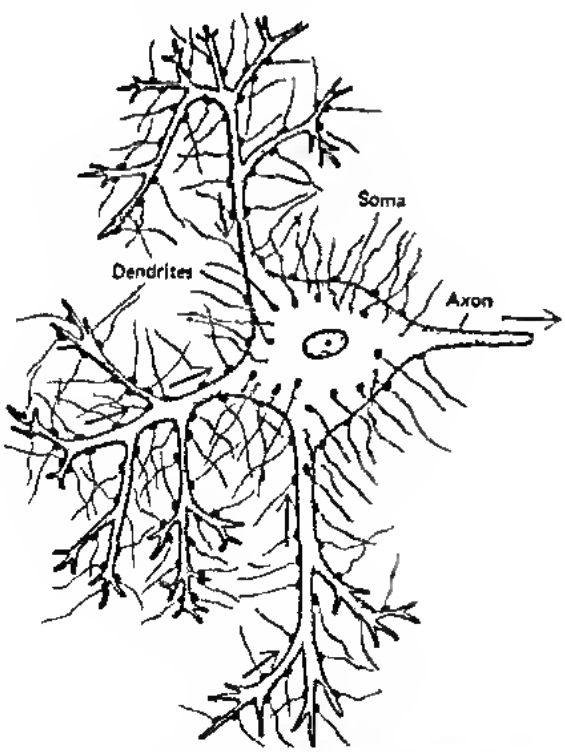
One-Way Conduction Through Chemical Synapses. Chemical synapses have one exceedingly important characteristic that makes them highly desirable as the form of transmission of nervous system signals: they always transmit the signals in one direction—that is, from the neuron that secretes the transmitter, called the *presynaptic neuron*, to the neuron on which the transmitter acts, called the *postsynaptic neuron*. This is the principle of *one-way conduction* through chemical synapses, and it is quite different from conduction through electrical synapses that can transmit signals in either direction.

Think for a moment about the extreme importance of the one-way conduction mechanism. It allows signals to be directed toward specific goals. Indeed, it is this specific transmission of signals to discrete and highly focused areas in the nervous system that allows the nervous system to perform its myriad functions of sensation, motor control, memory, and many others.

PHYSIOLOGIC ANATOMY OF THE SYNAPSE

Figure 46–4 illustrates a typical *motor neuron* in the anterior horn of the spinal cord. It is composed of three major parts: the *soma*, which is the main body of the neuron; a single *axon*, which extends from the soma into the peripheral nerve; and the *dendrites*, which are thin projections of the soma that extend up to 1 mm into the surrounding areas of the cord

An average of about 6000 small knobs called *presynaptic terminals* lie on the surfaces of the dendrites and soma of the motor neuron, approximately 80 to 90 per cent of them on the dendrites and only 10 to 20 per cent on the soma These terminals are the ends of nerve fibrils that originate in many other neurons, and usually not more than a few of the terminals are derived from any single previous neuron. Later it will become evident that many of these presynaptic terminals are *excitatory* and secrete a substance that excites the postsynaptic neuron, whereas others are *inhibitory* and secrete a substance that inhibits the neuron.



46–4. A typical motor neuron, showing presynaptic s on the neuronal soma and dendrites. Note also the axon.

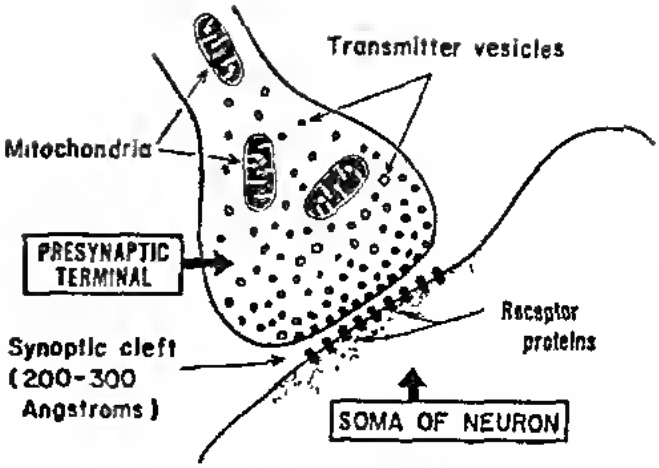


Figure 46–5. Physiologic anatomy of the synapse

Neurons in other parts of the cord and brain differ markedly from the motor neuron in (1) the size of the cell body, (2) the length, size, and number of dendrites, ranging in length from almost none at all up to as long as many centimeters, (3) the length and size of the axon, and (4) the number of presynaptic terminals, which may range from only a few to more than a hundred thousand. These differences make neurons in different parts of the nervous system react differently to incoming signals and therefore perform different functions

The Presynaptic Terminals. Electron microscopic studies of the presynaptic terminals show that they have varied anatomical forms, but most resemble small round or oval knobs and therefore are frequently called *terminal knobs, boutons, endfeet*, or *synaptic knobs*.

Figure 46–5 illustrates the basic structure of the presynaptic terminal. It is separated from the neuronal soma by a *synaptic cleft* having a width usually of 200 to 300 Angstroms. The terminal has two internal structures important to the excitatory or inhibitory functions of the synapse: the *synaptic vesicles* and the *mitochondria*. The synaptic vesicles contain a *transmitter substance* which, when released into the synaptic cleft, either *excites* or *inhibits* the neurons—excites if the neuronal membrane contains *excitatory receptors*, inhibits if it contains *inhibitory receptors*. The mitochondria provide ATP, which is required to synthesize new transmitter substance. The transmitter must be synthesized extremely rapidly because the amount stored in the vesicles is sufficient to last for only a few seconds to a few minutes of maximum activity.

When an action potential spreads over a presynaptic terminal, the membrane depolarization causes emptying of a small number of vesicles into the cleft, and the released transmitter in turn causes an immediate change in the permeability characteristics of the postsynaptic neuronal mem-

brane, which leads to excitation or inhibitioa of the neuron, depending on its receptor characteristics.

Mechaniam by Which Action Potentials Cause Transmitter Release at the Presynaptic Terminals—Role of Calcium Ions. The synaptic membrane of the presynaptic terminals contains large numbers of *voltage-gated calcium channels.* This is quite different from the other areas of the nerve fiber which contain very few of these channels. When the action potential depolarizes the terminal, large numbers of calcium ions, along with the sodium ions that cause the action potential, flow into the terminal. And the quantity of transmitter substance that is released into the synaptic cleft is directly related to the number of calcium ions that enter the terminal. The precise mechanism by which the calcium ions cause this release is not known but it is believed to be the following:

When the calcium ions enter tho synaptic terminal, it is believed that they bind with protein receptors on the inside surfaces of the eynaptic membrane called *release sites.* This in turn causes the transmitter vesicles in the local vicinity to bind with the membrene and actually to fuse with it, and finally to open to the exterior by the process called *exocytosis* that was described in Chapter 2. Often eeveral hundred vesicles release their transmitter into the cleft following e single action potential.

Beceuse of the requirement for calcium ions to cause release of the eynaptic transmitter, any conditioa that reducee the quantity of calcium ione entering the presynaptic terminal aiso will decrease the amount of transmitter released. Some conditions that can cause thie ere (a) a weak action potential caused by preliminery partial depolarization of the presynaptic terminal, (b) decrease in concentration of calcium ions in the extracellular fluid, or (c) any factor that reduces the permeability of the presynaptic terminal to calcium ions.

One transmitter substance thet occurs in certain parts of the nervous system is acetylcholine, as is discussed later. It has been calculated that about 3000 molecules of acetylcholine are present in each vesicle, and enough vesicles are present in the presynaptic terminal on a neuron to transmit a few thousand impulses.

Synthesia of New Transmitter Suhstance. Fortunately, the presynaptic terminals have the capability of continually synthesizing new transmitter substance. Were it not for this, synaptic transmission would become completely ineffective within a few minutes. The synthesis occurs either partially or totally in the cytoplasm of the presynaptic terminals, and then the newly synthesized transmitter is immediately absorbed into the vesicles and stored until needed.

As an example, acetylcholine is synthesized from acetyl-CoA and choline in the presence of the enzyme *choline acetyltransferase,* an enzyme that is present in abundance in the cytoplasm of the cholinergic type of presynaptic terminal. When acetylcholine is released from the terminal into the synaptic cleft, it is rapidly split again to acetate and choline by the enzyme *cholinesterase* that is adherent to the proteoglycon reticulum that fills the space of the synaptic cleft. Then the choline is actively transported back into the terminal to be used once more for synthesis of new acetylcholine.

Recycling of the Vesicles. Not only are some of the transmitter substances recycled, such as recycling of cholme, but the vesicles themselves are also recycled. When they fuse with the synaptic membrane and open up to the exterior, the extra membrane supplied by the vesicles at first simply enlarges the synaptic membrane. However, shortly thereafter, the vesicle portions of the synaptic membrane invagnate back toward the inside of the terminal and then punch off to form new vesicles. Furthermore, these new vesicles still contain the appropriate transport proteins that are required for concentrating new transmitter substance in the vesicles.

Thus the vesicles are used again and again. But even so, both the vesicles and the mitochondria that supply the energy for transmitter synthesis eventually disintegrate. Fortunately, new vesicles and mitochondria are continually transported from the cell soma down the axon to the presynaptic terminal, moving along the axon at a velocity of about 40 cm per day, thus replenishing the supply in the terminals.

Action of the Transmitter Substance on the Postsynaptic Neuron—The Function of Receptors. At the synapse, the membrane of the postsynaptic neuron conteins lerge numbers of *receptor proteins,* illustreted in Figure 46–5. These receptors have two important components: (1) a *binding component* thet protrudes outward from the membrane into the aynaptic cleft—it binds with the neurotransmitter from the presynaptic terminel, and (2) an *ionophore component* that protrudes through the membrene to the interior of the postsynaptic neuron The ionophore in turn is one of two types: (1) a *chemically activated ion channel,* or (2) an *enzyme that activates an internal metabolic system of the cell.*

The chemically activated ion channels (also called ligand-activated channels) are mainly of three types: (a) *sodium channels* that allow mainly sodium ions (but some potassium ions as well) to pass through, (b) *potassium channels* that allow mainly potassium ions to pass, and (c) *chloride channels* that allow chloride and a few other anions to pass. We shall learn later that opening the sodium channels excites the postsynaptic neuron. Therefore, a transmitter substance that opens the sodium channels is called an *excitatory transmitter.* Oa the other hand, opening of potassium and chloride channels inhibits the neuron, and transmitters that open either or both of these are called *inhibitory transmitters.*

Activation of an enzymatic type of receptor causes other effects on the postsynaptic neuron. One effect is to *activate cellular genes,* which in

turn manufacture additional receptors for the postsynaptic membrane. Another effect is to *activate protein kinases* that decrease the numbers of receptors. Changes such as these can alter the reactivity of the synapse for minutes, days, months, or even years. Therefore, transmitter substances that cause such effects are called synaptic *modulators* rather than excitatory or inhibitory transmitters. Recent experiments have demonstrated that such modulators are important in at least some of the memory processes, which we shall discuss in Chapter 54.

Duration of Action of the Transmitter, and Removal of the Transmitter from the Synapse. When either an excitatory or an inhibitory transmitter is released into the synaptic cleft, it binds almost instantly with its specific chemically activated ion channels and opens these either to the flow of sodium ions in the case of the excitatory transmitter or the flow of potassium or chloride ions in the case of the inhibitory transmitter. However, after these channels have remained open for only 1 to 2 milliseconds, in most instances, they close equally as rapidly. The reason for this is that the transmitter agent itself is rapidly removed from the synaptic cleft. This is achieved in three different ways.

(1) By *diffusion* of the transmitter out of the cleft into the surrounding fluids.

(2) By *enzymatic destruction* within the cleft itself. For instance, in the case of acetylcholine, the enzyme *cholinesterase* is present in the cleft, bound in the proteoglycan matrix that fills the space. Each molecule of this enzyme can split as many as ten molecules of acetylcholine each millisecond, thus inactivating this transmitter substance. Similar effects occur for other transmitters.

(3) Many of the transmitters are *actively transported back into the presynaptic terminal itself* and are then reused again and again. This is called *transmitter re-uptake*. It occurs especially prominently at the presynaptic terminals of the sympathetic nervous system for the re-uptake of norepinephrine, as we shall discuss in Chapter 57.

The degree to which each of these methods of removal is utilized is different for each type of transmitter.

CHEMICAL AND PHYSIOLOGICAL NATURES OF THE TRANSMITTER SUBSTANCES

Excitation and Inhibition. Whether a transmitter will cause excitation or inhibition is determined not only by the nature of the transmitter but also by the nature of the receptor in the postsynaptic membrane. To give an example, the same neuron might be excited at a synapse that releases acetylcholine but inhibited at still another synapse that releases glycine. Thus, the neuronal membrane at the first synapse contains an *excitatory receptor* for acetylcholine and at the second synapse an *inhibitory receptor* for glycine. To give another example, norepinephrine released at some synapses in the central nervous system causes inhibition whereas at other synapses it causes excitation. In the first case, the postsynaptic neuronal membranes contain an inhibitory receptor for norepinephrine while the others contain an excitatory receptor for the same transmitter.

Chemical Substances That Function As Neurotransmitters

More than 30 different chemical substances have either been proved or postulated to be synaptic transmitters; most are listed in Table 46–1. In general, there are four different classes of transmitter substances:

Class 1: Acetylcholine in a class by itself.
Class 2: Several different amines.
Class 3: Several different amino acids.
Class 4: Neuroactive peptides.

The different neurotransmitters have been identified or postulated in a number of different ways:

(1) by actually isolating the transmitter substance from the neuronal tissue;

(2) by showing that the appropriate enzymes are available for synthesis of the specific transmitter;

Table 46–1 NEUROTRANSMITTERS

Class I:	
	Acetylcholine
Class II: The Amines	
	Norepinephrine
	Epinephrine
	Dopamine
	Serotonin
Class III: Amino Acids	
	γ-Aminobutyric acid (GABA)
	Glycine
	Glutamate
Class IV: Peptides	
A	Hypothalamic-releasing hormones
	Thyrotropin-releasing hormone
	Luteinizing hormone–releasing hormone
	Somatostatin (growth hormone–inhibitory factor)
B	Pituitary peptides
	ACTH
	β Endorphin
	α-Melanocyte-stimulating hormone
	Vasopressin
	Oxytocin
C	Peptides that act on gut and brain
	Leucine enkephalin
	Methionine enkephalin
	Substance P
	Cholecystokinin
	Vasoactive intestinal polypeptide (VIP)
	Neurotensin
	Insulin
	Glucagon
D	From other tissues
	Angiotensin II
	Bradykinin
	Carnosine
	Bombesin

(3) by demonstrating that injection of the transmitter substance into local neuronal areas will cause excitation or inhibition;

(4) by demonstrating specific enzymes for removing the transmitter substance after it is released; and

(5) in several other ways.

Some of the more important of the transmitters are the following:

Acetylcholine is secreted by neurons in many areas of the brain, but specifically by the large pyramidal cells of the motor cortex, by many different neurons in the basal ganglia, by the motor neurons that innervate the skeletal muscles, by the preganglionic neurons of the autonomic nervous system, by the postganglionic neurons of the parasympathetic nervous system, and by some of the postganglionic neurons of the sympathetic nervous system. In most instances acetylcholine has an excitatory effect; however, it is known to have inhibitory effects at some of the peripheral parasympathetic nerve endings, such as inhibition of the heart by the vagus nerves.

Norepinephrine is secreted by many neurons whose cell bodies are located in the brain stem and hypothalamus. Specifically, norepinephrine-secreting neurons located in the *locus ceruleus* in the pons send nerve fibers to widespread areas of the brain and help control the overall activity and mood of the mind. In many of these areas it probably causes excitation, but in others inhibition. Norepinephrine is also secreted by most of the postganglionic neurons of the sympathetic nervous system, where it excites some organs but inhibits others.

Dopamine is secreted by neurons that originate in the substantia nigra. The terminations of these neurons are mainly in the striatal region of the basal ganglia. The effect of dopamine is usually inhibition.

Glycine is secreted mainly at synapses in the spinal cord. It probably always acts as an inhibitory transmitter.

Gamma-aminobutyric acid (GABA) is secreted by nerve terminals in the spinal cord, the cerebellum, the basal ganglia, and many areas of the cortex. It is believed always to cause inhibition.

Glutamate is probably secreted by the presynaptic terminals in many of the sensory pathways as well as in many areas of the cortex. It probably always causes excitation.

Substance P is probably released by pain fiber terminals in the dorsal horns of the spinal cord. And it is also found in the basal ganglia and hypothalamus. In general, it causes excitation.

Enkephalins are probably secreted by nerve terminals in the spinal cord, in the brain stem, in the thalamus, and in the hypothalamus. These probably act as excitatory transmitters to excite other systems that inhibit the transmission of pain.

Serotonin is secreted by nuclei that originate in the median raphe of the brain stem and project to many brain areas, especially to the dorsal horns of the spinal cord and to the hypothalamus. Serotonin acts as an inhibitor of pain pathways in the cord, and it is also believed to help control the mood of the person, perhaps even to cause sleep.

Despite this long list of transmitter substances, it is only a partial list. Other substances that have been proved or suggested include peptides, other amino acids, histamine, prostaglandins, cyclic AMP, and many others.

Release of Only a Single Type of Transmitter Substance by Each Neuron. Except in special cases, it is believed that each neuron releases only one type of transmitter, and it releases this same transmitter at all of its separate terminals. This is frequently called the *Dale principle* in honor of the physiologist most instrumental in developing this concept.

However, like all good theories, a number of exceptions have now been found in which the same nerve terminal might secrete a transmitter that acts as an excitatory or inhibitory transmitter and yet another that acts as a modulator. For instance the excitatory transmitter might be glutamate which causes immediate excitation, but at the same time one of the neuropeptides might be secreted as a modulator to increase the number of glutamate receptors, in this way increasing the sensitivity of the same synapse for days or weeks. Unfortunately, little is yet known about the modulator mechanisms. Therefore, for the most part, in the remainder of this chapter we will discuss principally the excitatory and inhibitory effects of the neurotransmitters.

ELECTRICAL EVENTS DURING NEURONAL EXCITATION

The electrical events in neuronal excitation have been studied in the large motor neurons of the anterior horn of the spinal cord. Therefore, the events to be described in the following few sections pertain essentially to these neurons. However, except for some quantitative differences, they apply to most other neurons of the nervous system as well.

The Resting Membrane Potential of the Neuronal Soma. Figure 46–6 illustrates the soma of a motor neuron, showing the resting membrane potential to be about −65 millivolts. This is somewhat less than the −80 to −90 millivolts found in large peripheral nerve fibers and in skeletal muscle fibers; the lower voltage is important, however, because it allows both positive and negative control of the degree of excitability of the neuron. That is, decreasing the voltage to a less negative value makes the membrane of the neuron more excitable, whereas increasing this voltage to a more negative value makes the neuron less excitable. This is the basis of the two modes of function of the neuron—either excitation or inhibition—as we will explain in detail in the following sections.

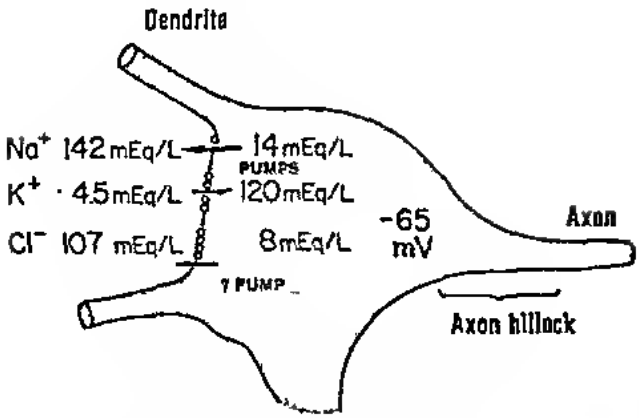


Figure 46–6. Distribution of sodium, potassium, and chloride ions across the neuronal somal membrane, origin of the intrasomal membrane potential

Concentration Differences of Ions Across the Neuronal Somal Membrane. Figure 46–6 also illustrates the concentration differences across tha neuronal somal membrane of the three ions that are most important for neuronal function· sodium ions, potassium ıons, and chloride ions

At the top, the sodium ion concentration ıs ahown to be very great in tha extracellular fluid but low inside tha neuron This sodium concentration gradient is caused by a strong sodium pump that continually pumps sodium out of the neuron.

The figure also shows that the potassium ion concentration ıs large inside the neuronal soma but very low ın the extracellular fluid. It illustrates that there Is also a potassium pump (the other half of the Na^+·K^+ pump, as described ın Chapter 9) that tends to pump potassium to the ınterior while there ıs a very high degree of permeability to potassium The potassium ıons leak through the neuronal somal pores so readily that this nullifies most but not all of the effectiveness of the pump.

Figure 46–6 shows the chloride ion to be of high concentration ın the extracellular fluid but low concentration ınside the neuron It also shows that the membrane is highly permeable to chloride ions and that there may be a weak chloride pump. But whether there is or is not a chloride pump, most of the reason for the low concentration of chloride ions inside the neuron ıs the −65 millivolts in the neuron. That is, this negative voltage repels the negatively charged chloride ıons, forcing them outward through the pores until the concentration difference is much greater outside the membrane than on the ınside.

Let us recall at this point what we learned in Chapters 9 and 10 about the relationship of ionic concentration differences to membrane potentials. It will be recalled that an electrical potential across the membrane can exactly oppose the movement of ions through a membrane despite concentration differences between the outside and inside of the membrane if the potential is of the proper [illegible] and magnitude. Such a potential that exactly opposes movement of each type of ion is called the Nernst potential; the equation for this is the following:

$$\text{EMF(mV)} = \pm 61 \times \log\left(\frac{\text{Concentration outside}}{\text{Concentration inside}}\right)$$

where EMF is the Nernst potential in millivolts on the *inside of the membrane* The potential will be positive (+) for a positive ion and negative (−) for a negative ion.

Now, let us calculate the Nernst potential that will exactly oppose the movement of each of the three separate ions: sodium, potassium, and chloride.

For the sodium concentration difference shown in Figure 46–6, 142 mEq/liter on the exterior and 14 mEq/liter on the ınterior, the membrane potential that would exactly oppose sodium Ion movement through the sodium channels would be +61 millivolts. However, the actual membrane potential is −65 millivolts, not +61 millivolts. Therefore, sodium ions normally diffusa ınward through the sodium channels; however, not many sodlum ions will diffuse beeause most of tha sodium channels are normally closed. Furthermore, those sodium ions that do diffusa to the interior are normally pumped immediately back to the extarior by the sodium pump.

For potassium ions, the concentration gradient is 120 mEq/liter inside the neuron and 4.5 mEq/liter outside Thls gives a Nernst potential of −86 millivolts insde the neuron, which Is more negative than the −65 that actually exists Therefore there is a tendency for potassium lons to diffuse to the outslde of the neuron, but this is opposed by the continual pumping of these potassium ıons back to the interior.

Finally, the chloride ion gradient, 107 mEq/liter outside and 8 mEq/liter inside, yields a Nernst potential of −70 millivolts inside the neuron, which ia slightly more negative than the actual value measured. Therefore, chloride ions tend normally to leak to the interior of the neuron, but those that do diffuse are moved back to the exterior, perhaps by an active chloride pump

Keep these three Nernst potentials in mind and also remember the direction in which the different ions tend to diffuse, for this information will be important in understanding both excitation and ınhibition of the neuron by synaptic transmission

Origin of the Resting Membrane Potential of the Neuronal Soma. The basic cause of the −65 millivolt resting membrane potential of the neuronal soma is the sodium-potassium pump This pump causes the extrusion of more positively charged sodium ions to the exterior than potassium to the interior—3 sodium ions for each 2 potassium ions. Since there are large numbers of negatively charged ions inside the soma that cannot diffuse through the membrane—protein ions, phosphate

ions, and many others—extrusion of the excess positive ions to the exterior leaves all these nondiffusible negative ions inside the cell unbalanced by positive ions. Therefore, the interior of the neuron becomes negatively charged as the result of the sodium-potassium pump. This principle was discussed in more detail in Chapter 10 in relation to the resting membrane potential of nerve fibers. In addition, as also explained in Chapter 10, diffusion of potassium ions outward through the membrane is another cause of intracellular negativity.

Uniform Distribution of the Potential Inside the Soma. The interior of the neuronal soma contains a very highly conductive electrolytic solution, the intracellular fluid of the neuron. Furthermore, the diameter of the neuronal soma is very large (from 10 to 80 microns in diameter), causing there to be almost no resistance to conduction of electrical current from one part of the somal interior to another part. Therefore, any change in potential in any part of the intrasomal fluid causes an almost exactly equal change in potential at all other points inside the soma. This is an important principle because it plays a major role in the summation of signals entering the neuron from multiple sources, as we shall see in subsequent sections of this chapter.

Effect of Synaptic Excitation on the Postsynaptic Membrane—The Excitatory Postsynaptic Potential. Figure 46–7A illustrates the resting neuron with an unexcited presynaptic terminal resting upon its surface. The resting membrane potential everywhere in the soma is −65 millivolts.

Figure 46–7B illustrates a presynaptic terminal that has secreted a transmitter into the cleft between the terminal and the neuronal somal membrane. This transmitter acts on a membrane excitatory receptor *to increase the membrane's permeability to both* Na^+ and K^+. However, because of the large electrochemical gradient that tends to move sodium inward and because the potassium concentration gradient is not far from equilibrium with the electrical potential, this large opening of the membrane pores mainly allows sodium ions to rush to the inside of the membrane.

The rapid influx of the positively charged sodium ions to the interior of the neuron neutralizes part of the negativity of the resting membrane potential. Thus, in Figure 46–7B the resting membrane potential has been increased from −65 millivolts to −45 millivolts. This increase in voltage above the normal resting neuronal potential—that is, to a less negative value—is called the *excitatory postsynaptic potential* (or EPSP) because when this potential rises high enough it will elicit an action potential in the neuron, thus exciting it. In this case the EPSP is + 20 millivolts.

However, we must issue a word of warning at this point. Discharge of a single presynaptic terminal can never increase the neuronal potential from −65 millivolts up to −45 millivolts. Instead, an increase of this magnitude requires the simultaneous discharge of many terminals—about 70 for the usual anterior motor neuron—at the same time or in rapid succession. This occurs by a process called *summation*, which will be discussed in detail in the following sections.

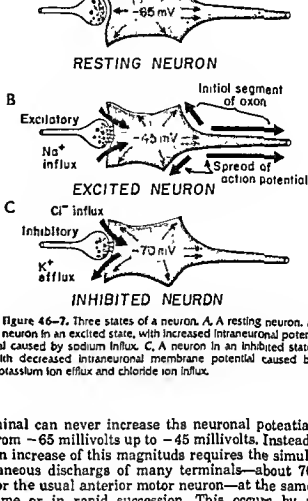


Figure 46–7. Three states of a neuron. *A*, A resting neuron. *B*, A neuron in an excited state, with increased intraneuronal potential caused by sodium influx. *C*, A neuron in an inhibited state, with decreased intraneuronal membrane potential caused by potassium ion efflux and chloride ion influx.

Generation of Action Potentials at the Axon Hillock of the Neuron—Threshold for Excitation. When the membrane potential inside the neuron rises high enough, there comes a point at which this initiates an action potential in the neuron. However, the action potential does not begin on the somal membrane adjacent to the excitatory synapses. Instead, it begins in the *axon hillock*, which is the point of origin of the axon from the neuronal soma. The main reason for this point of origin of the action potential is that the soma has relatively few voltage-gated sodium channels in its membrane, which makes it difficult to open the required number of channels to elicit an action potential. On the other hand, the membrane of the axon hillock has seven times as great a concentration of voltage-gated sodium channels and therefore can generate an action potential with much greater ease than can the soma. The excitatory postsynaptic potential that will elicit an action potential at the axon hillock is between +15 and +20 millivolts. This is in contrast to approximately +30 millivolts required on the soma.

Once the action potential begins, it travels both peripherally along the axon and also backward over the soma. In many instances, it travels backward into the dendrites, too, but not into all of them because they, like the neuronal soma, also have very few voltage-gated sodium channels.

Thus, in Figure 46–7B, it is shown that under normal conditions the *threshold* for excitation of the neuron is about −45 millivolts, which represents an excitatory postsynaptic potential of +20 millivolts—that is, 20 millivolts more positive than the normal resting neuronal potential of −65 millivolts.

ELECTRICAL EVENTS IN NEURONAL INHIBITION

Effect of Inhibitory Synapses on the Postsynaptic Membrane—The Inhibitory Postsynaptic Potential. It was pointed out earlier that the excitatory synapses open "sodium" channels in the postsynaptic membrane, and these allow easy passage mainly of sodium ions. And, because of the very large difference between the resting membrane potential and the Nernst potential for sodium, especially large numbers of sodium ions move into the neuron and cause the excitatory postsynaptic potential. The inhibitory synapses, by contrast, open the potassium and chloride channels, allowing easy passage of both these ions. Now, to understand how the inhibitory synapses inhibit the postsynaptic neuron, we must recall what we learned about the Nernst potentials for both the potassium ions and the chloride ions. We calculated this potential for potassium ions to be about −86 millivolts and for chloride ions about −70 millivolts. Both these potentials are more negative than the −65 millivolts normally present inside the resting neuronal membrane. Therefore, opening the potassium and the chloride channels will allow potassium ions to move to the exterior, which will make the membrane potential more negative than normal and will allow the chloride ions to move to the interior, which also will make the membrane potential more negative than usual. This increases the degree of intracellular negativity, which is called *hyperpolarization*. It obviously inhibits the neuron because the membrane potential is now farther away than ever from the threshold for excitation. Therefore, the increase in negativity beyond the normal resting membrane potential level is called the *inhibitory postsynaptic potential* (IPSP).

Thus, Figure 46–7C illustrates the effect on the membrane potential caused by excitation of inhibitory synapses, allowing chloride influx into the cell and potassium efflux from the cell, with the membrane potential decreasing from its normal value of −65 millivolts to the more negative value of −70 millivolts. This membrane potential that is 5 millivolts more negative is the inhibitory postsynaptic potential. Thus the IPSP in this instance is −5 millivolts.

Inhibition of Neurons Without Causing an Inhibitory Postsynaptic Potential—"Short Circuiting" of the Membrane. Sometimes activation of the inhibitory synapses causes little or no inhibitory postsynaptic potential but nevertheless still inhibits the neuron.

The reason that the potential often does not change is that in some neurons the concentration differences across the membrane for the potassium and chloride ions are only able to cause a diffusion potential equal to the normal resting potential. Therefore, when the inhibitory pores open, there is no net flow of ions to cause an inhibitory postsynaptic potential. Yet, both the potassium and the chloride ions do diffuse bidirectionally through the wide-open pores many times as rapidly as normally, and this high flux of these two ions inhibits the neuron in the following way: When excitatory synapses fire and sodium ions flow into the neuron, this now causes far less excitatory postsynaptic potential than usual because any tendency for the membrane potential to change away from the resting potential is immediately opposed by rapid flux of potassium and chloride ions through the inhibitory pores to bring the potential back to the negative equilibrium potential for these two ions. Therefore, the influx of sodium ions required to cause excitation may be as much as 5 to 20 times normal.

This tendency for the potassium and chloride ions to maintain the membrane potential near the resting value when the inhibitory pores are wide open is called "short circuiting" of the membrane, thus making the sodium current flow caused by excitatory synapses ineffective in exciting the cell.

To express the phenomenon of short circuiting more mathematically, one needs to recall the Goldman equation from Chapter 10. This equation shows that the membrane potential is determined by summation of the tendencies for the different ions to carry electrical charges through the membrane in the two directions. The membrane potential will approach the Nernst equilibrium potential for those ions that permeate the membrane to the greatest extent. When the inhibitory channels are wide open, the chloride and potassium ions permeate the membrane greatly. Therefore, when the excitatory channels open, it is difficult to raise the neuronal potential up to the threshold value for excitation.

Presynaptic Inhibition

In addition to the inhibition caused by inhibitory synapses operating at the neuronal membrane, called *postsynaptic inhibition*, another type of inhibition often occurs before the signal reaches the synapse. This type of inhibition, called *presynaptic inhibition*, is believed to occur in the following way.

In presynaptic inhibition, the inhibition is caused by "presynaptic" synapses that lie on the terminal nerve fibrils before they themselves terminate on the following neuron. It is believed that

these presynaptic synapses secrete a transmitter substance that in some way not yet understood depresses the voltage of the action potential that occurs at the synaptic membrane of the terminal and, as has already been pointed out, this greatly decreases the amount of calcium ions that enter the terminal and therefore also the amount of transmitter released by the terminal. Therefore, the degree of excitation of the neuron is also greatly suppressed, or inhibited.

Presynaptic inhibition occurs in many of the sensory pathways in the nervous system. That is, the adjacent nerve fibers inhibit each other, which minimizes the spread of signals from one fiber to the next. We will discuss this phenomenon more fully in the following chapter.

The precise mechanism by which the presynaptic synapses function is not known. However, since most of the nerve fibers involved are excitatory types of fibers, it has been suggested that presynaptic inhibition might be caused by excitatory synapses that cause partial depolarization of the terminal nervo fibrils on which they lie. This in turn would reduce the intensity of the action potential passing into this terminal fibril and thereby reduce the release of transmitter substance because even a slight reduction of the magnitude of the action potential reduces calcium entry into the presynaptic terminal and this reduction causes an even far greater reduction in the amount of transmitter released.

Presynaptic inhibition is different from postsynaptic inhibition in its time sequence. It requires many milliseconds to develop, but once it does occur, it can last for aa long as minutes or even hours. Postsynaptic inhibition, at least of the anterior motor neurons, lasts for only 10 to 15 milliseconds.

SUMMATION OF POSTSYNAPTIC POTENTIALS

Time Course of Postsynaptie Potentials. When a synapse excites the anterior motor neuron the neuronal membrane becomes highly permeable for only 1 to 2 milliseconds. During this time sodium ions diffuse rapidly to the interior of the cell to increase the intraneuronal potential, thus creating the *excitatory postsynaptic potential*. This potential then persists for about 15 milliseconds, because this is the time required for the positive charges to flow into the dendrites and axon or for potassium ions to leak out or chloride ions to leak in to re-establish the normal resting membrane potential.

Precisely the opposite effect occurs for the inhibitory postsynaptic potential. That is, the inhibitory synapse increases the permeability of the membrane to potassium and chloride ions for 1 to 2 milliseconds, and this usually decreases the intraneuronal potential to a more negative value than normal, thereby creating the *inhibitory postsynaptic potential*. This potential also persists for about 15 milliseconds.

However, other types of transmitter substances acting on other neurons can perhaps excite or inhibit for hundreds of milliseconds or even for seconds, minutes, or hours.

Spatial Summation of the Postsynaptie Potentials. It has already been pointed out that excitation of a single presynaptic terminal on the surface of a neuron will almost never excite the neuron. The reason for this is that sufficient transmitter substance is released by a single terminal to cause an excitatory postsynaptic potential usually no more than a millivolt at most, instead of the required 15 to 20 millivolts to reach the usual threshold for excitation. However, during excitation in a neuronal pool of the nervous system, many presynaptic terminals are usually stimulated at the same time, and even though these terminals are spread over wide areas of the neuron, their effects can still summate. The reason for this summation is the following: It has already been pointed out that a change in the potential at any single point within the soma will cause the potential to change everywhere in the soma almost exactly equally. Therefore, for each excitatory synapse that discharges simultaneously, the intrasomal potential becomes more positive by as much as a small fraction of a milllivolt up to about 1 millivolt. When the excitatory postsynaptic potential becomes great enough, the threshold for firing will be reached, and an action potential will generate at the axon hillock. This effect is illustrated in Figure 46–8, which shows several excitatory postsynaptic potentials. The bottom postsynaptic potential in the figure was caused by stimulation of only four excitatory synapses; then the next higher potential was caused by stimulation of two times as many synapses; finally, a still higher excitatory postsynaptic potential was caused by stimulation of four times as many synapses. This time an action potential was generated at the axon hillock.

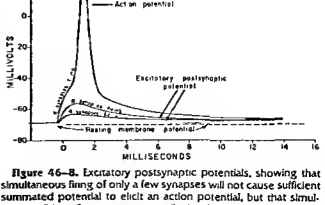


Figure 46–8. Excitatory postsynaptic potentials, showing that simultaneous firing of only a few synapses will not cause sufficient summated potential to elicit an action potential, but that simultaneous firing of many synapses will raise the summated potential to the threshold for excitation and cause a superimposed action potential.

This effect of summing simultaneous postsynaptic potentials by excitation of multiple terminals on widely spaced areas of the membrane is called *spatial summation.*

Temporal Summation. Most presynaptic terminals can fire repetitively in rapid succession only a few milliseconds apart. Each time a terminal fires, the released transmitter substance opens the membrane channels for a millisecond or so. Since the postsynaptic potential lasts up to 15 milliseconds, a second opening of the same channel can increase the postsynaptic potential to a still greater level so that the more rapid the rate of terminal stimulation, the greater the effective postsynaptic potential. Thus, successive postsynaptic potentials of individual presynaptic terminals, if they occur rapidly enough, can summate in the same way that postsynaptic potentials can summate from widely distributed terminals over the surface of the neuron. This summation is called *temporal summation.*

Simultaneous Summation of Inhibitory and Excitatory Postsynaptic Potentials. Obviously, if an inhibitory postsynaptic potential is tending to decrease the membrane potential to a more negative value while an excitatory postsynaptic potential is tending to increase the potential at the same time, these two effects can either completely nullify each other or partially nullify each other. Also, inhibitory "short circuiting" of the membrane potential can nullify much of an excitatory potential. Thus, if a neuron is currently being excited by an excitatory postsynaptic potential, then an inhibitory signal from another source can easily reduce the postsynaptic potential to less than the threshold value for excitation, thus turning off the activity of the neuron.

Facilitation of Neurons. Often the summated postsynaptic potential is excitatory in nature but has not risen high enough to reach the threshold for excitation. When this happens the neuron is said to be *facilitated.* That is, its membrane potential is nearer the threshold for firing than normally but not yet to the firing level. Nevertheless, a signal entering the neuron from some other source can then excite the neuron very easily. Diffuse signals in the nervous system often facilitate large groups of neurons so that they can respond quickly and easily to signals arriving from second sources.

SPECIAL FUNCTIONS OF DENDRITES IN EXCITING NEURONS

The Large Spatial Field of Excitation of the Dendrites. The dendrites of the anterior motor neurons extend for 0.5 to 1 millimeter in all directions from the neuronal soma. Therefore, these dendrites can receive signals from a large spatial area around the motor neuron. This provides vast opportunity for summation of signals from many separate presynaptic neurons.

It is also important that between 80 and 90 per cent of all the presynaptic terminals terminate on the dendrites of the anterior motor neuron in contrast to only 10 to 20 per cent terminating on the neuronal soma. Therefore, the preponderant share of the excitation of the neuron is provided by signals transmitted over the dendrites.

Failure of Many Dendrites to Transmit Action Potentials. Many dendrites fail to transmit action potentials because their membranes have relatively few voltage-gated sodium channels, so that their thresholds for excitation are very high. Yet they do transmit *electrotonic current* down the dendrites to the soma. (Transmission of electrotonic current means the direct spread of current by electrical conduction in the fluids of the dendrites with no generation of action potentials.) Stimulation of the neuron by this current has special characteristics, as follows:

Decrement of Electrotonic Conduction in the Dendrites—Greater Excitation by Synapses Near the Soma. In Figure 46–9 a number of excitatory and inhibitory synapses are shown stimulating the dendrites of a neuron. On the two dendrites to the left in the figure are shown excitatory effects near the ends of the dendrites; note the high levels of the excitatory postsynaptic potentials at these ends—that is, the less negative membrane potentials. However, a large share of the excitatory postsynaptic potential is lost before it reaches the soma. The reason for this is that the dendrites are long and thin, and their membranes are also leaky to electrical current. Therefore, before the excitatory potentials can reach the soma, a large share of the potential is lost by leakage through the membrane. This decrease in membrane potential as it spreads electrotonically along dendrites toward the soma is called *decremental conduction.*

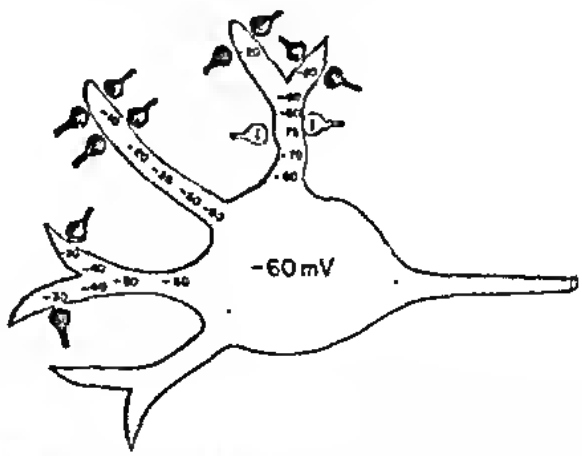


Figure 46–9. Stimulation of a neuron by presynaptic terminals located on dendrites, showing, especially, decremental conduction of excitatory electrotonic potentials in the two dendrites to the left and inhibition of dendritic excitation in the dendrite that is uppermost.

It is also obvious that the nearer the excitatory synapse is to the soma of the neuron, the less will be the decrement of conduction. Therefore, those synapses that lie near the soma have far more excitatory effect than those that lie far away from the soma.

Rapid Re-Excitation of the Neuron by the Dendrites After the Neuron Fires. When an action potential is generated in a neuron, this action potential spreads back over the soma but not always over the dendrites. Therefore, the excitatory postsynaptic potentials in the dendrites often are only partially disturbed by the action potential, so that just as soon as the action potential is over, the potentials still existing in the dendrites are ready and waiting to excite the neuron again. Thus, the dendrites have a "holding capacity" for the excitatory signal from presynaptic sources.

Summation of Excitation and Inhibition in Dendrites. The uppermost dendrite of Figure 46–9 is shown to be stimulated by both excitatory and inhibitory synapses. At the tip of the dendrite is a strong excitatory postsynaptic potential, but nearer to the soma are two inhibitory synapses acting on the same dendrite. These inhibitory synapses provide a hyperpolarizing voltage that completely nullifies the excitatory effect and indeed transmits a small amount of inhibition by electrotonic conduction toward the soma. Thus, dendrites can summate excitatory and inhibitory postsynaptic potentials in the same way that the soma can.

RELATION OF STATE OF EXCITATION OF THE NEURON TO THE RATE OF FIRING

The "Excitatory State." The *"excitatory state"* of a neuron is defined as the degree of excitatory drive to the neuron. If there is a higher degree of excitation than inhibition of the neuron at any given instant, then it is said that there is an *excitatory state*. On the other hand, if there is more inhibition than excitation, then it is said that there is an *inhibitory state*.

When the excitatory state of a neuron rises above the threshold for excitation, then the neuron will fire repetitively as long as the excitatory state remains at this level. However, *the rate at which it will fire is determined by how much* the excitatory state is above threshold. To explain this, we must first consider what happens to the neuronal somal potential during and following the action potential.

Changes in Neuronal Somal Potential During and Following the Action Potential. Figure 46–10 illustrates an action potential spreading backward over the neuronal soma after being initiated at the axon hillock by an excitatory postsynaptic potential. Following the spike portion of the action potential, there is a very long *postive after-potential*—that is, a state of "hypolarization"—lasting for many milliseconds. During this interval the somal membrane potential falls below the normal resting membrane potential of ~65 millivolts. This is probably caused by a high degree of permeability of the neuronal membrane to potassium ions that persists for many milliseconds after the action potential is over. And, in addition, the high membrane conductivity for potassium ions also "short circuits" excitatory potentials.

The importance of this state of hyperpolarization and short circuiting after the spike potential is that the neuron remains in an *inhibited state* during this period of time. Therefore, a far greater excitatory state is required during this time than normally to cause re-excitation of the neuron.

Relationship of Excitatory State to Frequency of Firing. The curve shown at the top of Figure 46–10, labeled "Excitatory state required for re-excitation," depicts the relative level of the excitatory state required at each instant after an action potential is over to re-excite the neuron. Note that very soon after an action potential is over, a very high excitatory state is required. That is, a very large number of excitatory synapses must be firing simultaneously. Then, after many milliseconds have passed and the state of hyperpolarization and the short circuiting of the neuron have begun to disappear, the excitatory state required becomes greatly reduced.

Therefore, it is immediately evident that when the excitatory state is high, a second action potential will appear very soon after the previous one. Then still a third action potential will appear soon after the second, and this process will continue indefinitely. Thus, at a very high excitatory state the frequency of firing of the neuron is great.

On the other hand, when the excitatory state is only barely above threshold, the neuron must recover almost completely from the hyperpolarization and short circuiting, which requires many

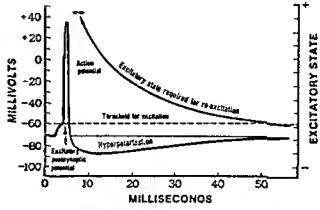


Figure 46–10. A neuronal potential followed by a prolonged period of neuronal hyperpolarization. Also shown is the "excitatory state" required for re-excitation of the neuron at given intervals after the action potential is over.

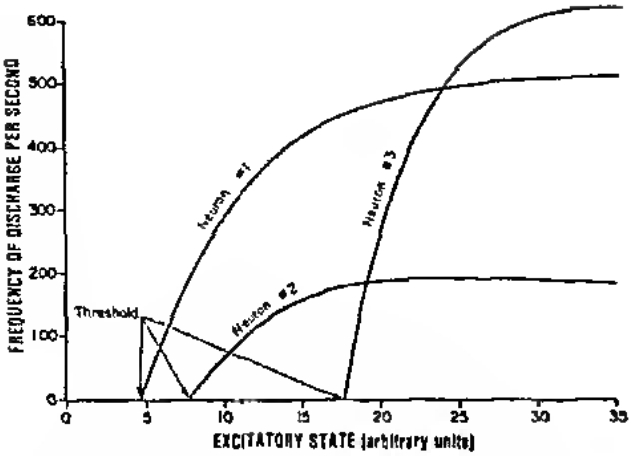


Figure 46–11. Response characteristics of different types of neurons to progressively increasing levels of excitatory state

milliseconds, before it will fire again. Therefore, the frequency of neuronal firing is low.

Response Characteristics of Different Neurons to Increasing Levels of Excitatory State. Histological study of the nervous system immediately convinces one of the widely varying types of neurons in different parts of the nervous system. And, physiologically, the different types of neurons perform different functions. Therefore, as would be expected, the ability to respond to stimulation by the synapses varies from one type of neuron to another.

Figure 46–11 illustrates theoretical responses of three different types of neurons to varying levels of excitatory state. Note that neuron #1 has a low threshold for excitation while neuron #3 has a high threshold But note also that neuron #2 has the lowest maximum frequency of discharge, while neuron #3 has the highest maximum frequency.

Some neurons in the central nervous system fire continuously because even the normal excitatory state is above the threshold level. Their frequency of firing can usually be increased still more by further increasing their excitatory state Or, the frequency may be decreased, or firing even be stopped, by superimposing an inhibitory state on the neuron

Thus, different neurons respond differently, have different thresholds for excitation, and have widely differing maximal frequencies of discharge. With a little imagination one can readily understand the importance of having neurons with many different types of response characteristics to perform the widely varying functions of the nervous system

SOME SPECIAL CHARACTERISTICS OF SYNAPTIC TRANSMISSION

Fatigue of Synaptic Transmission. When excitatory synapses are repetitively stimulated at a rapid rate, the number of discharges by the postsynaptic neuron is at first very great, but it becomes progressively less in succeeding milliseconds or seconds. This is called *fatigue* of synaptic transmission.

Fatigue is an exceedingly important characteristic of synaptic function, for when areas of the nervous system become overexcited, fatigue causes them to lose this excess excitability after a while. For example, fatigue is probably the most important means by which the excess excitability of the brain during an epileptic convulsion is finally subdued so that the convulsion ceases. Thus, the development of fatigue is a protective mechanism against excess neuronal activity. This will be discussed further in the description of reverberating neuronal circuits in the following chapter.

The mechanism of fatigue is mainly exhaustion of the stores of transmitter substance in the synaptic terminals, particularly since it has been calculated that the excitatory terminals can store enough excitatory transmitter for only 10,000 normal synaptic transmissions, an amount that can be exhausted in only a few seconds to a few minutes. However, part of the fatigue process probably also results from two other factors as well: (1) progressive inactivation of many of the postsynaptic membrane receptors, and (2) slow buildup of calcium ions inside the postsynaptic neuronal cell caused by the successive action potentials—these calcium ions in turn open calcium-activated potassium channels which cause an inhibitory effect on the postsynaptic neuron.

Post-Tetanic Facilitation. When a rapidly repetitive series of impulses stimulates an excitatory synapse for a period of time and then a rest period is allowed, the postsynaptic neuron will usually be even more responsive to subsequent stimulation than normally. This is called *post-tetanic facilitation.*

Experiments have shown that post-tetanic facilitation is caused mainly by the buildup of excess calcium ions in the presynaptic terminals, caused by the successive influx of more calcium ions with each action potential because the calcium pump pumps too slowly to remove all of these immediately. These accumulated calcium ions add to the effects of the action potentials themselves in causing more and more vesicular release of transmitter substance Occasionally the effect can be great enough to cause the release of transmitter at a rate two times normal

The physiological significance of post-tetanic facilitation is still very doubtful, and it may have no real significance at all. However, since post-tetanic facilitation can last from a few seconds in some neurons to several hours in others, it is immediately apparent that neurons could possibly store information by this mechanism Therefore post-tetanic facilitation might well be a mechanism of "short-term" memory in the central nervous system. This possibility will be discussed further in Chapter 54 in relation to the memory function of the cerebral cortex.

Effect of Acidosis and Alkalosis on Synaptic Transmission. The neurons are highly responsive to changes in pH of the surrounding interstitial fluids *Alkalosis greatly increases neuronal excitability* For in-

stance, a rise in arterial pH from the normal of 7.4 to about 7.8 often causes cerebral convulsions because of increased excitability of the neurons. This can be demonstrated especially well by having a person who is predisposed to epileptic convulsions overbreathe. The overbreathing elevates the pH of the blood only momentarily, but even this short interval can often precipitate an epileptic attack.

On the other hand, *acidosis greatly depresses neuronal activity;* a fall in pH from 7.4 to below 7.0 usually causes a comatose state. For instance, in very severe diabetic or uremic acidosis, coma always develops.

Effect of Hypoxia on Synaptic Transmission. Neuronal excitability is also highly dependent on an adequate supply of oxygen. Cessation of oxygen supply for only a few seconds can cause complete inexcitability of the neurons. This is often seen when the cerebral circulation is temporarily interrupted, for within 3 to 5 seconds the person becomes unconscious.

Effect of Drugs on Synaptic Transmission. Many different drugs are known to increase the excitability of neurons, and others are known to decrease the excitability. For instance, caffeine, theophylline, and theobromine, which are found in coffee, tea, and cocoa, respectively, all increase neuronal excitability, presumably by reducing the threshold for excitation of the neurons. However, strychnine, which is one of the best known of all the agents that increase the excitability of neurons, does not reduce the threshold for excitation of the neurons at all but, instead, *inhibits the action of at least some of the inhibitory transmitters* on the neurons, probably especially the inhibitory effect of glycine in the spinal cord. In consequence, the effects of the excitatory transmitters become overwhelming, and the neurons become so excited that they go into rapidly repetitive discharge, resulting in severe convulsions.

Most anesthetics increase the membrane threshold for excitation and thereby decrease synaptic transmission at many points in the nervous system. Because most of the anesthetics are lipid-soluble, it has been reasoned that they might change the physical characteristics of the neuronal membranes, making them less responsive to excitatory agents.

Synaptic Delay. In transmission of an action potential from a presynaptic neuron to a postsynaptic neuron, a certain amount of time is consumed in the process of (a) discharge of the transmitter substance by the presynaptic terminal, (b) diffusion of the transmitter to the postsynaptic neuronal membrane, (c) action of the transmitter on the membrane receptor, (d) action of the receptor to increase the membrane permeability, and (e) inward diffusion of sodium to raise the excitatory postsynaptic potential to a high enough value to elicit an action potential. The *minimal* period of time required for all these events to take place, even when large numbers of presynaptic terminals are stimulated simultaneously, is about 0.5 millisecond. This is called the *synaptic delay.* It is important for the following reason: Neurophysiologists can measure the *minimal* delay time between an input volley of impulses and an output volley and from this can estimate the number of series neurons in the circuit.

REFERENCES

Addink, A D (ed) Exogenous and Endogenous Influences on Metabolic and Neural Control New York, Pergamon Press, 1982

Barde, Y A, et al New neurotrophic factors. *Annu Rev Physiol*, *45* 601, 1983

Berg, D K New neuronal growth factors. *Annu Rev Neurosci*, *7* 149, 1984

Buchtel, H A The Conceptual Nervous System New York, Pergamon Press, 1982

Changeux, J P, *et al* Acetylcholine receptor An allosteric protein *Science*, *225* 1335, 1984

Creese, I, *et al* The classification of dopamine receptors Relationship to radioligand binding *Annu Rev Neurosci* 6 43, 1983

Eccles, J C My scientific odyssey *Annu Rev Physiol*, *39* 1, 1977

Eyzaguirre, C. Physiology of the Nervous System Chicago, Year Book Medical Publishers, 1965

Fernstrom, J D Dietary precursors and brain neurotransmitter formation. *Annu Rev Med*, *32* 413, 1981

Ganong, W F The brain renin angiotensin system *Annu Rev Physiol*, *46* 17, 1984

Hanin, I (ed.) Dynamics of Neurotransmitter Function New York, Raven Press, 1984

Iversen, L L. Nonopioid neuropeptides in mammalian CNS *Annu Rev Pharmacol Toxicol* *23* 1, 1983

Kehoe, J, and Marty, A Certain slow synaptic responses. Their properties and possible underlying mechanisms. *Annu. Rev Biophys Bioeng* 9 437, 1980

Kennedy, M B Experimental approaches to understanding the role of protein phosphorylation in the regulation of neuronal function *Annu Rev Neurosci*, *6* 493, 1983

Kostyuk, P G Intracellular perfusion of nerve cells and its effects on membrane currents. *Physiol Rev*, *64* 435, 1984

Krnjevic, K Transmitters in motor systems *In* Brooks, V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 107

Laduron, P M Presynaptic heteroreceptors in regulation of neuronal transmission. *Biochem Pharmacol*, *34* 487, 1985

Landis, S C Neuronal growth cones. *Annu Rev Physiol*, *45* 567, 1983

Loh, H H, and Law, P Y The role of membrane lipids in receptor mechanisms *Annu Rev Pharmacol Toxicol*, *20* 201, 1980

Lynch, G, and Schubert, P The use of in vitro brain slices for multidisciplinary studies of synaptic function. *Annu Rev Neurosci*, *3* 1, 1980

McKay, R D G Molecular approach to the nervous system *Annu Rev Neurosci*, *6* 527, 1983

McKerns, K W (ed) Hormonally Active Brain Peptides Structure and Function. New York, Plenum Publishing Corp, 1982

Nicoll, R A, et al Substance P as a transmitter candidate *Annu Rev Neurosci*, *3* 227, 1980

Popot, J-L, and Changeux, J-P Nicotinic receptor of acetylcholine: Structure of an oligomeric integral membrane protein *Physiol Rev*, *64* 1162, 1984

Purves, D, and Lichtman, J W. Specific connections between nerve cells *Annu. Rev Physiol*, *45* 553, 1983

Reichardt, L F Immunological approaches to the nervous system. *Science*, *225* 1294, 1984

Schubert, D. Developmental Biology of Cultured Nerve, Muscle and Glia New York, John Wiley & Sons, 1984

Silverman, A-J Magnocellular neurosecretory system *Annu Rev Neurosci.*, *6* 357, 1983

Stein, J F Introduction to Neurophysiology St. Louis, C V Mosby, 1982

Su, C. Purinergic neurotransmission and neuromodulation. *Annu. Rev Pharmacol Toxicol*, *23* 397, 1983

Tucek, S. Regulation of acetylcholine synthesis in the brain *J Neurochem.*, *44* 11, 1985

Wright, E M: Electrophysiology of plasma membrane vesicles. *Am J Physiol*, *246* F363, 1984.

47

Neuronal Mechanisms and Circuits for Processing Information

Certain types of neuronal mechanisms and neuronal circuits are used again and again throughout the nervous system. The purpose of this chapter is to characterize some of these and to preview some of their important uses.

INFORMATION, SIGNALS, AND IMPULSES

The term *information*, as it applies to the nervous system, means a variety of different things, such as knowledge, facts, quantitative values, intensity of pain, intensity of light, temperature, and any other aspect of the body or its immediate surroundings that has meaning. Thus, pain from a pin prick is information, pressure on the bottom of the feet is information, degree of angulation of the joints is information, and a stored memory in the brain is information

However, information cannot be transmitted in its original form but only in the form of action potentials, also called *nerve impulses*. Thus, a part of the body that is subjected to pain must first convert this information into nerve impulses; specific areas of the brain convert abstract thoughts also into nerve impulses that are then transmitted either elsewhere in the brain or into peripheral nerves to motor effectors throughout the body. The retina of the eye converts vision into nerve impulses, the nerve endings of the joints convert degree of angulation of the joints into nerve impulses, and so forth

Signals. In the transmission of information, it is frequently not desirable to speak in terms of the individual impulses but instead of the overall pattern of impulses; this pattern is called a *signal* As an example, when pressure is applied to a large area of skin, impulses are transmitted by large numbers of parallel nerve fibers, and the total pattern of impulses transmitted by all these fibers is a signal. Thus, we can speak of visual signals, auditory signals, somesthetic sensory signals, motor signals, and so forth

TRANSMISSION OF SIGNALS IN NERVE TRACTS

SIGNAL STRENGTH

One of the characteristics of information that always must be conveyed is its quantitative intensity, for instance the intensity of pain. The different gradations of intensity can be transmitted either by utilizing increasing numbers of parallel fibers or by sending more impulses along a single fiber. These two mechanisms are called respectively, spatial summation and temporal summation

Spatial Summation (Multiple Fiber Summation). Figure 47–1 illustrates the phenomenon of spatial summation, whereby increasing signal strength is transmitted by using progressively greater numbers of fibers. This figure shows a section of skin innervated by a large number of parallel pain nerve fibers. Each of these arborizes into hundreds of minute *free nerve endings* that serve as pain receptors. The entire cluster of fibers from one pain fiber frequently covers an area of skin as large as 5 cm in diameter, and this area is called the *receptive field* of that fiber. The number of endings is large in the center of the field but diminishes toward the periphery. One can also see from the figure that the arborizing nerve fibrils overlap those from other pain fibers. Therefore, a pin prick of the skin usually stimulates endings from many different pain fibers simultaneously.

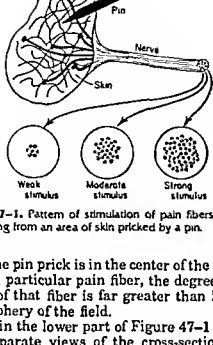


Figure 47–1. Pattern of stimulation of pain fibers in a nerve trunk leading from an area of skin pricked by a pin.

But, if the pin prick is in the center of the receptive field of a particular pain fiber, the degree of stimulation of that fiber is far greater than if it is in the periphery of the field.

Thus, in the lower part of Figure 47–1 is shown three separate views of the cross-section of the nerve bundle leading from the skin area. To the left is shown the effect of a weak stimulus, with only a single nerve fiber in the middle of the bundle stimulated very strongly (represented by the solid fiber) whereas several adjacent fibers are stimulated weakly (half-solid fibers). The other two views of the nerve cross-section show the effect respectively of a moderate stimulus and a strong stimulus, with progressively more fibers being stimulated. Thus, the stronger signals spread to more and more fibers. This phenomenon is called *spatial summation*, which means simply that one of the means by which signals of increasing strength are transmitted in the nervous system is by utilizing progressively greater numbers of fibers.

Temporal Summation. A second means for transmitting signals of increasing strength is by increasing the *frequency* of nerve impulses in each fiber, which is called *temporal summation*. Figure 47–2 illustrates this type of summation, showing in the upper part a changing strength of signal and in the lower part the actual impulses transmitted by the nerve fiber.

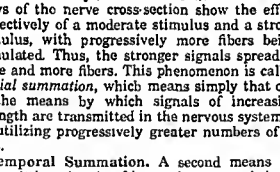


Figure 47–2. Translation of signal strength into a frequency-modulated series of nerve impulses, showing above the strength of signal and below the separate nerve impulses.

TRANSMISSION AND PROCESSING OF SIGNALS IN NEURONAL POOLS

The central nervous system is made up of literally hundreds or even thousands of separate neuronal pools, some of which contain very few neurons while others hold vast numbers. For instance, the entire cerebral cortex could be considered to be a single large neuronal pool. It has many separate fiber tracts coming to it (afferent fibers) and others leaving it (efferent fibers). Furthermore it maintains the same quality of spatial orientation as that found in the nerve bundles, individual points of the cortex connecting with specific points elsewhere in the nervous system or connecting through the peripheral nerves with specific points in the body. However, within this pool of neurons are large numbers of short nerve fibers whereby signals spread horizontally from neuron to neuron within the pool itself.

Other neuronal pools include the different basal ganglia and the specific nuclei in the thalamus, cerebellum, mesencephalon, pons, and medulla. Also, the entire dorsal gray matter of the spinal cord could be considered to be one long pool of neurons, and the entire anterior gray matter another long neuronal pool. Each pool has its own special characteristics of organization which cause it to process signals in its own special way, thus allowing these special characteristics to achieve the multitude of functions of the nervous system. Yet, despite their differences in function, the pools also have many similarities that are described in the following pages.

RELAYING OF SIGNALS THROUGH NEURONAL POOLS

Organization of Neurons for Relaying Signals. Figure 47–3 is a schematic diagram of several neurons in a neuronal pool, showing "input" fibers to the left and "output" fibers to the right. Each input fiber divides hundreds to thousands of times, providing an average of a thousand or more terminal fibrils that spread over a large area in the pool to synapse with the dendrites or cell bodies of the neurons in the pool. The dendrites of some of the neurons also arborize and spread in the pool. The neuronal area stimulated by each incoming nerve fiber is called its *stimulatory field*. Note that each input fiber arborizes so that large numbers of its terminals lie on the centermost neurons in its "field," but progressively fewer terminals lie on the neurons farther from the center of the field.

Threshold and Subthreshold Stimuli—Facilitation. From the discussion of synaptic function

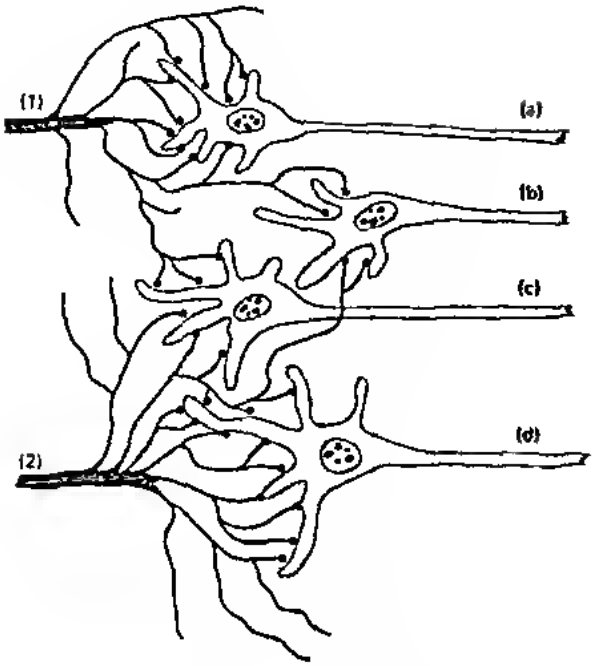


Figure 47–3. Basic organization of a neuronal pool

in the previous chapter, it will be recalled that discharge of a single excitatory presynaptic terminal almost never stimulates the postsynaptic neuron. Instead, large numbers of terminals must discharge on the same neuron either simultaneously or in rapid succession to cause excitation. For instance, in Figure 47–3, let us assume that six separate terminals must discharge simultaneously to excite any one of the neurons If the student will count the number of terminals on each one of the neurons from each input fiber, he or she will see that input fiber (1) has more than enough terminals to cause neuron (a) to discharge Therefore, the stimulus from input fiber (1) to this neuron is said to be an *excitatory stimulus*, or it is also called a *threshold stimulus* because it is above the threshold required for excitation.

Input fiber (1) also contributes terminals to neurons (b) and (c) but not enough to cause excitation. Nevertheless, discharge of these terminals makes both these neurons more excitable to signals arriving through other incoming nerve fibers Therefore, the stimulus to these neurons is said to be *subthreshold*, and the neurons are said to be *facilitated*

Similarly for input fiber (2) the stimulus to neuron (d) is a threshold stimulus and to neurons (b) and (c) a subthreshold, but facilitating, stimulus

It must be recognized that Figure 47–3 represents a highly condensed version of a neuronal pool, for each input nerve fiber usually provides terminals to hundreds or thousands of separate neurons in its distribution "field," as illustrated in Figure 47–4 In the central portion of the field, almost all the neurons are stimulated by the in-

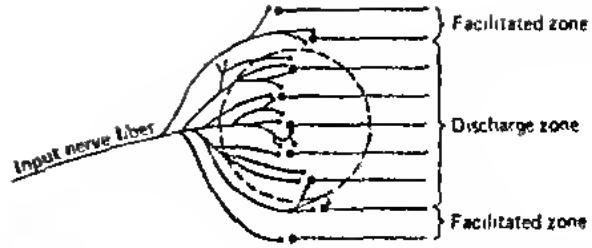


Figure 47–4. "Discharge" and "facilitated" zones of a neuronal pool

coming fiber, designated in Figure 47–4 by the darkened circle area Therefore, this is said to be the *discharge zone* of the incoming fiber, also called *excited zone* or *liminal zone*. To either side, the neurons are facilitated but not excited, and these areas are called the *facilitated zone*, or also called *subthreshold zone* or *subliminal zone*.

Inhibition of a Neuronal Pool. We must also remember that some incoming fibers inhibit neurons rather than excite them. This is exactly the opposite of facilitation, and the entire field of the inhibitory branches is called the *inhibitory zone*. However, the degree of inhibition in the center of this zone is very great because of large numbers of endings in the center while it becomes progressively less toward its edges.

Divergence of Signals Passing Through Neuronal Pools

Often it is important for signals entering a neuronal pool to excite far greater numbers of nerve fibers leaving the pool. This phenomenon is called *divergence*. Two major types of divergence occur and have entirely different purposes. These are illustrated in Figure 47–5 and may be described as follows

An *amplifying* type of divergence often occurs, illustrated in Figure 47–5A. This means simply that an input signal spreads to an increasing number of neurons as it passes through successive

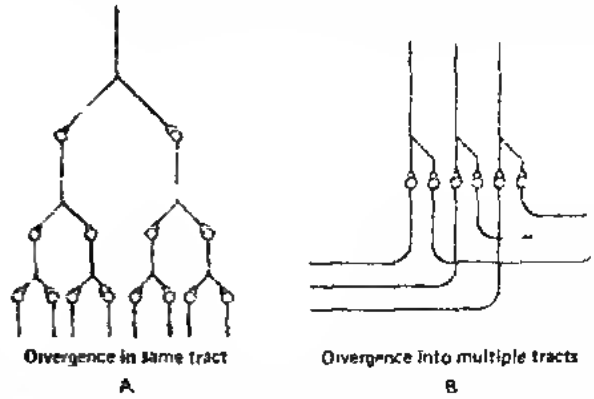


Figure 47–5. "Divergence" in neuronal pathways. *A*, Divergence within a pathway to cause "amplification" of the signal *B*, Divergence into multiple tracts to transmit the signal to separate areas

pools of a nervous pathway. This type of divergence is characteristic of the corticospinal pathway in its control of skeletal muscles, as follows: Stimulation of a single large pyramidal cell in the motor cortex transmits a single impulse into the spinal cord. Yet, under appropriate conditions of very strong cord facilitation, this impulse can stimulate perhaps several hundred interneurons and anterior motor neurons. Each of the motor neurons then stimulates as many as 100 or more muscle fibers. Thus, there is a total divergence, or amplification, of as much as 10,000-fold.

The second type of divergence, illustrated in Figure 47–5B, is *divergence into multiple tracts*. In this case, the signal is transmitted in two separate directions from the pool, to different parts of the nervous system where it is needed. For instance, information transmitted in the dorsal columns of the spinal cord takes two courses in the lower part of the brain: (1) into the cerebellum, and (2) on through the lower regions of the brain to the thalamus and cerebral cortex. Likewise, in the thalamus almost all sensory information is relayed both into deep structures of the thalamus and to discrete regions of the cerebral cortex.

Convergence of Signals

Often it is important for signals from multiple incoming neurons to excite the same neuron. That is, these incoming fibers *converge* on this same neuron. Like divergence, there are also two basic types of convergence, illustrated in Figure 47–6.

Section A of this figure shows *convergence from a single source*. That is, multiple terminals from an incoming fiber tract terminate on the same neuron. The importance of this is that neurons are almost never excited by an action potential from a single input terminal. But action potentials from multiple input terminals will provide enough "summation" to bring the neuron to the threshold required for discharge.

However, *convergence can also result from input signals* (excitatory or inhibitory) *from several different sources*, which is illustrated in Figure 47–6B. For instance, the interneurons of the spinal cord receive converging signals from (a) peripheral nerve fibers entering the cord, (b) propriospinal fibers passing from one segment of the cord to another, (c) corticospinal fibers from the cerebral cortex, and (d) several other long pathways descending from the brain into the spinal cord. Then the signals from the interneurons converge on the anterior motor neurons to control muscle function.

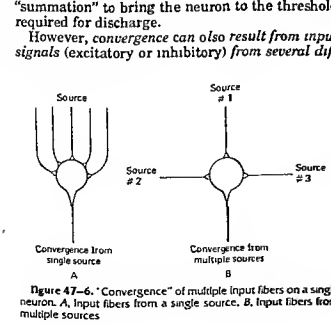


Figure 47–6. "Convergence" of multiple input fibers on a single neuron. *A*, Input fibers from a single source. *B*, Input fibers from multiple sources

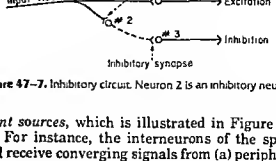


Figure 47–7. Inhibitory circuit. Neuron 2 is an inhibitory neuron.

Such convergence allows summation of information from different sources, and the resulting response is a summated effect of all the different types of information. Obviously, therefore, convergence is one of the important means by which the central nervous system correlates, summates, and sorts different types of information.

Neuronal Circuit Causing Both Excitatory and Inhibitory Output Signals

Sometimes an incoming signal to a neuronal pool causes an output excitatory signal going in one direction and at the same time an inhibitory signal going elsewhere. For instance, at the same time that an excitatory signal is transmitted by one set of neurons in the spinal cord to cause forward movement of a leg, an inhibitory signal is transmitted simultaneously through a separate set of neurons to inhibit the muscles on the back of the leg so that they will not oppose the forward movement. This type of circuit is characteristic of control of all antagonistic pairs of muscles, and it is called the *reciprocal inhibition circuit*.

Figure 47–7 illustrates the means by which the inhibition is achieved. The input fiber directly excites the excitatory output pathway, but it stimulates an intermediate *inhibitory neuron* (neuron 2) which then inhibits the second output pathway from the pool. This type of circuit is probably also important in preventing overactivity in many parts of the brain.

TRANSMISSION OF SPATIAL PATTERNS THROUGH SUCCESSIVE NEURONAL POOLS

Most information is transmitted from one part of the nervous system to another through several successive neuronal pools. For instance, sensory information from the skin passes first through the peripheral nerve fibers, then through second order neurons that originate either in the spinal cord or

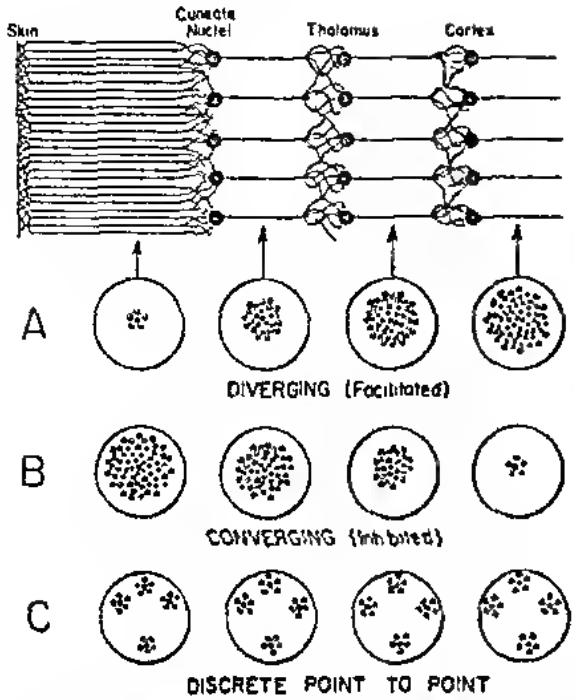


Figure 47–8. Typical organization of a sensory pathway from the skin to the cerebral cortex. *Below:* The patterns of fiber stimulation at different points in the pathway following stimulation by a pinprick when the pathway is *A*, facilitated, *B* inhibited, and *C*, normally excitable

in the cuneate and gracile nuclei of the medulla, and finally through third order neurons originating in the thalamus to the cerebral cortex. Such a pathway is illustrated at the top of Figure 47-8. Note that the sensory nerve endings in the skin overlap each other tremendously; and the terminal fibrils of each nerve fiber, on entering each neuronal pool, spread to many adjacent neurons, innervating perhaps 100 or more separate neurons. On first thought, one would expect signals from the skin to become completely mixed up by this haphazard arrangement of terminal fibrils in each neuronal pool. For statistical reasons, however, this does not occur, which can be explained as follows:

First, if a single point is stimulated in the skin, the nerve fiber with the most nerve endings in that particular spot becomes stimulated to the strongest extent, while the immediately adjacent nerve fibers become stimulated less strongly, and the nerve fibers still farther away become stimulated only weakly. When this signal arrives at the first neuronal pool, the stimulus spreads in many directions in the terminal fibrils of the neuronal pool. Yet the *greatest number of excited presynaptic terminals* lies very near the point of entry of the most excited input nerve fiber. Therefore, the neuron closest to this central point is the one that becomes stimulated to the greatest extent. Exactly the same effect occurs in the second neuronal pool in the thalamus and again when the signal reaches the cerebral cortex.

Yet, it is true that a signal passing through *highly facilitated neuronal pools* could diverge so much that the spatial pattern at the terminus of the pathway would be completely obscured. This effect is illustrated in Figure 47–8A, which shows successively expanding spatial patterns of neuron stimulation in such a facilitated, diverging pathway.

However, the degree of facilitation of the different neuronal pools varies from time to time. Under some conditions the degree of facilitation is so low that the pathway becomes converging, as illustrated in Figure 47–8B. In this case, a broad area of the skin is stimulated, but the signal loses part of its fringe stimuli as it passes through each successive pool until the breadth of the stimulus becomes contracted at the opposite end. One can achieve this type of stimulation by pressing ever so lightly with a flat object on the skin. The signal converges to give the person a sensation of almost a point contact.

In Figure 47–8C, four separate points are simultaneously stimulated on the skin, and the degree of excitability in each neuronal pool is exactly that amount required to prevent either divergence or convergence. Therefore, a reasonably true spatial pattern of each of the four points of stimulation is transmitted through the entire pathway.

Centrifugal Control of Neuronal Facilitation in the Sensory Pathways. It is obvious from the preceding discussion that the degree of facilitation of each neuronal pool must be maintained at exactly the proper level if faithful transmission of the spatial pattern is to occur. Recent discoveries have demonstrated that the degrees of inhibition or facilitation of most—indeed, probably all—neuronal pools in the different sensory pathways are controlled by *centrifugal nerve fibers* that pass from the respective sensory areas of the cortex downward to the separate sensory relay neuronal pools. Thus, these nerve fibers undoubtedly help to control the faithfulness of signal transmission.

Lateral Inhibition to Provide Contrast in the Spatial Pattern. When a single point of the skin or other sensory area is stimulated, not only is a single fiber excited but a number of "fringe" fibers are excited less strongly at the same time, as already explained. Therefore, the spatial pattern is blurred even before the signal begins to be transmitted through the pathway. However, in almost all pathways that transmit very exact information, such as the critical sensory pathways, lateral *inhibitory circuits* inhibit the fringe neurons and re-establish a truer spatial pattern.

Figure 47–9A illustrates one type of lateral inhibitory circuit, showing that the nerve fibers of a pathway give off collateral fibers that excite inhibitory neurons. These inhibitory neurons in turn inhibit the less excited fringe neurons in the signal pathway. The effect of this on transmission of the spatial pattern is illustrated in Figure 47–9B,

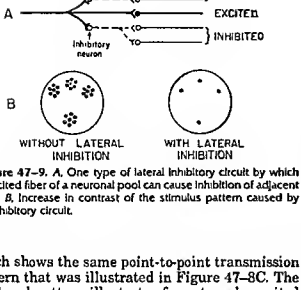


Figure 47–9. *A*. One type of lateral inhibitory circuit by which an excited fiber of a neuronal pool can cause inhibition of adjacent fibers. *B*. Increase in contrast of the stimulus pattern caused by the inhibitory circuit.

which shows the same point-to-point transmission pattern that was illustrated in Figure 47–8C. The left-hand pattern illustrates four strongly excited fibers (the solid fibers) with penumbras of fringe excitation surrounding each of these. In the illustration to the right, the penumbras have been removed by the lateral inhibitory circuits; obviously, this increases the *contrast* in the signal and helps in the *faithfulness* of transmission of the spatial pattern.

Unfortunately, we still know little about all the capabilities of the lateral inhibition mechanism, but the more we learn the more it becomes clear that this is one of the most useful of all the neuronal circuits of the nervous system to differentiate and analyze sensory signals. We shall discuss this much more fully in the following chapters.

PROLONGATION OF A SIGNAL BY A NEURONAL POOL—"AFTER-DISCHARGE"

Thus far, we have considered signals that are merely relayed through neuronal pools. However, in many instances, a signal entering a pool causes a prolonged output discharge, called *after-discharge*, even after the incoming signal is over, and lasting from a few milliseconds to as long as many minutes. The three basic mechanisms by which after-discharge occurs are the following:

Synaptic After-Discharge. When excitatory synapses discharge on the surfaces of dendrites or the soma of a neuron, a postsynaptic potential develops in the neuron and lasts for many milliseconds—in the anterior motor neuron for up to 15 milliseconds, though perhaps much longer in other neurons—especially so when some of the long-acting synaptic transmitter substances are involved. As long as this potential lasts it can continue to excite the neuron, causing it to transmit a continuous train of output impulses, as was explained in the previous chapter. Thus, as a result of this synaptic after-discharge mechanism alone, it is possible for a single instantaneous input to cause a sustained signal output (a series of repetitive discharges) lasting as long as 15 or more milliseconds.

The Parallel Circuit Type of After-Discharge. Figure 47–10 illustrates a second type of neuronal circuit that can cause intermediate periods of after-discharge. In this case, the input signal spreads through a series of neurons in the neuronal pool, and from many of these neurons impulses keep converging on an output neuron. It will be recalled that a signal is delayed at each synapse for about 0.5 millisecond, which is called the *synaptic delay*. Therefore, signals passing through a succession of intermediate neurons reach the output neuron one by one after varying periods of delay. Therefore, the output neuron continues to be stimulated for many milliseconds.

It is doubtful that more than a few dozen successive neurons ordinarily enter into a parallel after-discharge circuit. Therefore, one would suspect that this type of after-discharge circuit could cause after-discharges that last for no more than perhaps 25 to 50 milliseconds. Yet, this circuit does represent a means by which a single input signal, lasting less than 1 millisecond, can be converted into a sustained output signal lasting for a rather precise period of time.

The Reverberatory (Oscillatory) Circuit as a Cause of After-Discharge. Many neurophysiologists believe that one of the most important of all circuits in the entire nervous system is the *reverberatory*, or *oscillatory, circuit*. Such circuits are caused by positive feedback within the neuronal pool. That is, the output of a neuronal circuit feeds back to re-excite the same circuit. Consequently, once stimulated, the circuit discharges repetitively for a long time.

Several different postulated varieties of reverberatory circuits are illustrated in Figure 47–11, the simplest—in Figure 47–11A—involving only a single neuron. In this case, the output neuron simply sends a collateral nerve fiber back to its own dendrites or soma to restimulate itself; therefore, once the neuron should discharge, the feedback stimuli could theoretically help keep the neuron discharging for a long time thereafter.

Figure 47–11B illustrates a few additional neurons in the feedback circuit, which would give a longer period of time between the initial discharge and the feedback signal. Figure 47–11C illustrates

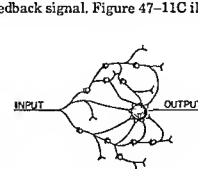


Figure 47–10. The parallel after-discharge circuit.

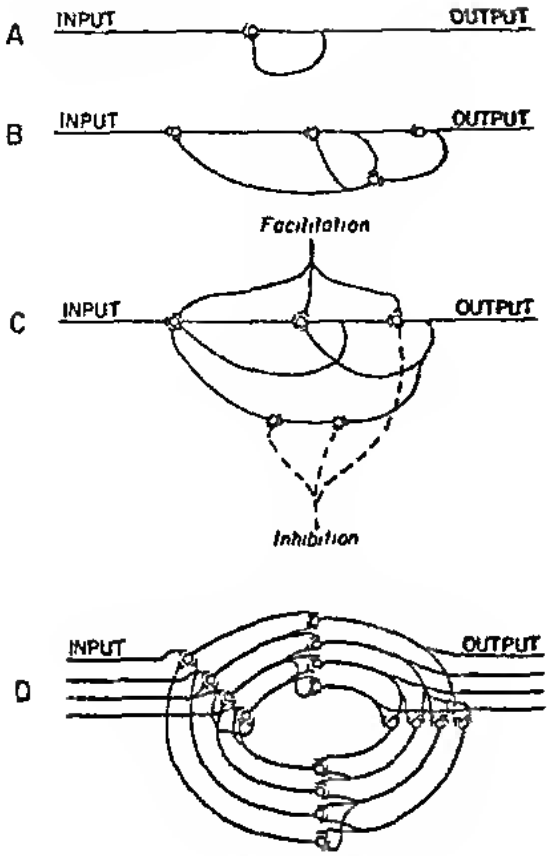


Figure 47–11. Reverberatory circuits of increasing complexity

a still more complex system in which both facilitatory and inhibitory fibers impinge on the reverberating pool. A facilitatory signal increases the ease with which reverberation takes place, whereas an inhibitory signal decreases the ease of reverberation.

Figure 47–11D illustrates that most reverberating pathways are constituted of many parallel fibers, and at each cell station the terminal fibrils diffuse widely. In such a system the total reverberating signal can be either weak or strong, depending on how many parallel nerve fibers are momentarily involved in the reverberation.

Finally, reverberation need not occur only in a single neuronal pool, for it can occur through a circuit involving two or more successive pools in the positive feedback pathway.

Characteristics of After-Discharge from a Reverberatory Circuit. Figure 47–12 illustrates postulated output signals from a reverberatory after-discharge circuit. The input stimulus need last only 1 millisecond or so, and yet the output can last for many milliseconds or even minutes. The figure demonstrates that the intensity of the output signal increases to a reasonably high value early in the reverberation, then decreases to a critical point, and at that point suddenly ceases entirely. Furthermore, the duration of the after-discharge is determined by the degree of inhibition or facilitation of the neuronal pool. In this way, signals from other parts of the brain can control the reaction of the pool to the input stimulus.

Almost these exact patterns of output signals can be recorded from the motor nerves exciting a muscle involved in the flexor reflex (discussed in Chapter 51), which is believed to be caused by a reverberating type of after-discharge following stimulation of pain fibers.

Importance of Synaptic Fatigue in Determining the Duration of Reverberation. It was pointed out in the previous chapter that synapses fatigue if stimulated for prolonged periods of time. Therefore, one of the most important factors that determines the duration of the reverberatory type of after-discharge is undoubtedly the rapidity with which the involved synapses fatigue. Rapid fatigue would obviously tend to shorten the period of after-discharge.

On the other hand, the greater the number of successive stages of neurons in the reverberatory pathway and the greater the number of collateral feedback fibrils, the easier would it be to keep the reverberation going.

Some Reverberatory Systems of the Nervous System and Their Durations of Reverberation. Even though neurophysiologists are not certain which functions of the nervous system can rightfully be ascribed to reverberatory circuits, some of those that have been postulated have reverberatory durations as short as 10 milliseconds or as long as several minutes, or perhaps even hours. Some examples include: (1) In an animal whose spinal cord is transected in the neck, a sudden painful stimulus to the animal's paw will cause the flexor muscles to contract and remain contracted from a fraction of a second to as much as several seconds after the stimulus ends. It is believed that this is caused by an after-discharge of the reverberatory type. (2) During respiration the inspiratory neuronal pool in the medulla becomes excited for about 2 seconds during each respiratory cycle. One theory suggests that this is caused by reverberation within the inspiratory neuronal pool. And (3) one theory of wakefulness is that continual reverberation occurs somewhere within the brain stem to keep a wakefulness area excited during the waking hours. If this be true,

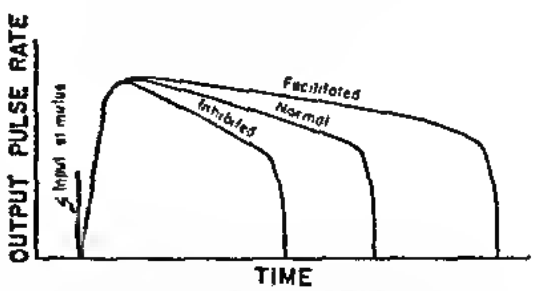


Figure 47–12. Typical pattern of the output signal from a reverberatory circuit following a single input stimulus, showing the effects of facilitation and inhibition

then this would represent a 14 to 18 hour period of reverberation.

Continuous Signal Output From Neuronal Pools

Some neuronal pools emit output signals continuously even without excitatory input signals. At least two different mechanisms theoretically can cause this effect: (1) intrinsic neuronal discharge and (2) reverberatory signals.

Continuous Discharge Caused by Intrinsic Neuronal Excitability. Neurons, like other excitable tissues, discharge repetitively if their membrane potentials rise above certain threshold levels. The membrane potentials of many neurons even normally are high enough to cause them to emit impulses continually. This occurs especially in large numbers of the neurons of the cerebellum as well as in most of the interneurons of the spinal cord. The rates at which these cells emit impulses can be increased by facilitatory signals or decreased by inhibitory signals; the latter can sometimes decrease the rate to extinction.

Continuous Signals Emitted from Reverberating Circuits as a Means for Transmitting Information. Obviously, a reverberating circuit that never fatigues to extinction could also be a source of continual impulses. Facilitatory impulses entering the reverberating pool, as illustrated in Figure 47–11C, could increase the output signal, and inhibition could decrease or even extinguish the output signal.

Figure 47–13 illustrates a continual output signal from a pool of neurons, whether it be a pool emitting impulses because of intrinsic neuronal excitability or as a result of reverberation. Note that an excitatory (or facilitatory) input signal greatly increases the output signal, whereas an inhibitory input signal greatly decreases the output. Those students who are familiar with radio transmitters will recognize this to be a *carrier wave* type of information transmission. That is, the excitatory and inhibitory control signals are not the *cause* of the output signal, but they do *control* it. Note that this carrier wave system

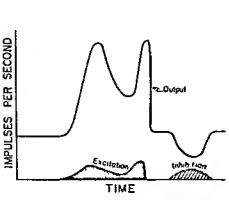


Figure 47–13. Continuous output from either a reverberating circuit or from a pool of intrinsically discharging neurons. This figure also shows the effect of excitatory or inhibitory input signals.

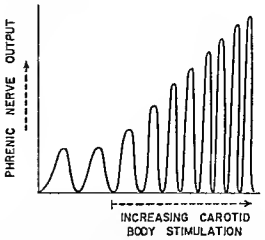


Figure 47–14. The rhythmical output from the respiratory center, showing that progressively increasing stimulation of the carotid body increases both the intensity and frequency of oscillation.

allows decrease in signal intensity as well as increase, whereas, up to this point, the types of information transmission that we have discussed have been only positive information rather than negative information. This type of information transmission is used by the autonomic nervous system to control such functions as vascular tone, gut tone, degree of constriction of the iris, heart rate, and others.

RHYTHMIC SIGNAL OUTPUT

Many neuronal circuits emit rhythmic output signals—for instance, the rhythmic respiratory signal originating in the reticular substance of the medulla and pons. This repetitive rhythmic signal continues throughout life, while other rhythmic signals, such as those that cause scratching movements by the hind leg of a dog or the walking movements in an animal, require input stimuli into the respective circuits to initiate the signals.

Many rhythmic signals are postulated to result from reverberating circuits or successive reverberating circuits that feed excitatory or inhibitory signals from one to the next. However, many other more complex mechanisms could also be responsible for rhythmic signals.

Obviously, facilitatory or inhibitory signals can affect rhythmic signal output in the same way that they can affect continuous signal outputs. Figure 47–14, for instance, illustrates the rhythmic respiratory signal in the phrenic nerve. However, when the carotid body is stimulated by arterial oxygen deficiency, the frequency and amplitude of the rhythmic signal pattern increase progressively.

INSTABILITY AND STABILITY OF NEURONAL CIRCUITS

Almost every part of the brain connects either directly or indirectly with every other part, and

this creates a serious problem If the first part excites the second, the second the third, the third the fourth, and so on until finally the signal re-excites the first part, it is clear that an excitatory signal entering any part of the brain would set off a continuous cycle of re-excitation of all parts. If this should occur, the brain would be inundated by a mass of uncontrolled reverberating signals—signals that would be transmitting no information but, nevertheless, would be consuming the circuits of the brain so that none of the informational signals could be transmitted Such an effect actually occurs in widespread areas of the brain during *epileptic convulsions*.

How does the central nervous system prevent this from happening all the time? The answer seems to lie in two basic mechanisms that function throughout the central nervous system: (1) inhibitory circuits, and (2) fatigue of synapses.

INHIBITORY CIRCUITS AS A MECHANISM FOR STABILIZING NERVOUS SYSTEM FUNCTION

The phenomenon of lateral inhibition that was discussed earlier in the chapter prevents signals in an informational pathway from spreading diffusely everywhere. In addition, two other types of inhibitory circuits in widespread areas of the brain help prevent excessive spread: (a) inhibitory feedback circuits that return from the terminal of pathways back to the initial excitatory neurons of the same pathways—these are believed to occur in all the sensory nervous pathways, and they inhibit the input neurons when the terminal become overly excited, and (b) some neuronal pools that exert gross inhibitory control over widespread areas of the brain—for instance, many of the basal ganglia exert inhibitory influences throughout the motor control system.

SYNAPTIC FATIGUE AS A MEANS FOR STABILIZING THE NERVOUS SYSTEM

Synaptic fatigue means simply that the signal becomes progressively weaker the more prolonged the period of excitation Figure 47–15 illustrates three successive records of a flexor reflex elicited in an animal caused by inflicting pain in the footpad of the paw. Note in each record that the strength of contraction progressively "decrements"—that is, its strength diminishes; this is believed to be caused by *fatigue* of the synapses in the flexor reflex circuit. Furthermore, the shorter the interval between the successive flexor reflexes, the less becomes the intensity of the reflex response. Thus, in most neuronal circuits that are overused, the sensitivities of the circuits become depressed

Automatic Short-Term Adjustment of Pathway Sensitivity by the Fatigue Mechanism. Now let us apply this phenomenon of fatigue to multiple pathways in the brain. Those that are

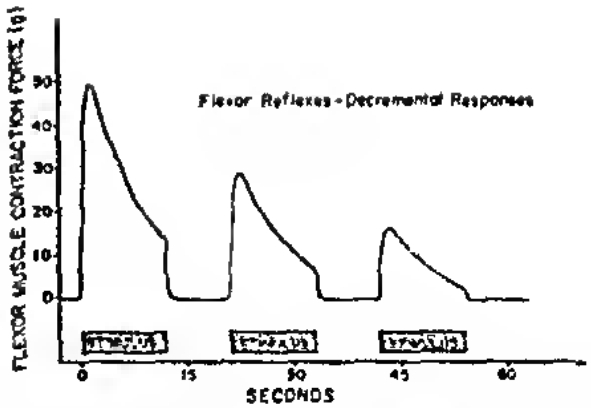


Figure 47–15. Successive flexor reflexes illustrating fatigue of conduction through the reflex pathway

overused usually become fatigued so that their sensitivities will be reduced. On the other hand, those that are underused will become rested and their sensitivities will increase. Thus, fatigue and recovery from fatigue constitute an important short-term means for moderating the sensitivities of the different nervous system circuits, helping to keep them operating in a range of sensitivity that allows effective function.

Long-Term Changes in Synaptic Sensitivity Caused by Automatic Downgrading or Upgrading of Synaptic Receptors. Recently it has been learned that the long-term sensitivities of synapses can be changed tremendously by downgrading the number of receptor proteins at the synaptic sites when there is overactivity and upgrading the receptors when there is underactivity. The mechanism for this is believed to be the following: Receptor proteins are being formed constantly by the endoplasmic reticular–Golgi apparatus system and are constantly being inserted into the synaptic membrane. However, when the synapses are overused and excesses of transmitter substance combine with the receptor proteins, many of these proteins are inactivated permanently and presumably removed from the synaptic membrane. This is especially true when some of the "modulator" transmitter substances are released at the synapses.

Therefore, in many if not most cases, overusage of a circuit will lead to gradually decreasing sensitivity of the synapses because of decreased receptor proteins, while underusage will cause prolonged increase in sensitivity. This is still another mechanism for continually adjusting the sensitivities of the respective neuronal circuits. It is indeed fortunate that fatigue and downgrading or upgrading of receptors, as well as other control mechanisms of the nervous system, continually adjust the sensitivity in each circuit to almost the exact level required for proper function. Think for a moment how serious it would be if the sensitivities of only a few of these circuits should be abnormally high; one might then expect almost continual muscle cramps, convulsions, psychotic disturbances,

hallucinations, tension, or many other nervous disorders. But the automatic controls normally readjust the sensitivities of the circuits back to a controllable range of reactivity any time the circuits begin to be too active.

REFERENCES

American Physiological Society Sensory Processes Washington, D C, American Physiological Society, 1984

An der Heiden, U Analysis of Neural Networks New York, Springer-Verlag, 1980

Anderson, H., *et al* Developmental neurobiology of invertebrates *Annu Rev. Neurosci*, 3 97, 1980

Asanuma, H, and Wilson, V J (eds) Integration in the Nervous System. New York, Igaku-Shoin, 1979

Baldissera, F., Hultborn, H, and Illert, M Integration in spinal neuronal systems. *In* Brooks, V B (ed) Handbook of Physiology Sec 1 Vol II Bethesda, American Physiological Society, 1981, p 509

Bennett, M V L. (ed) Synaptic Transmission and Neuronal Interaction New York, Raven Press, 1975

Björklund, A., and Stenevi, U Regeneration of monoaminergic and cholinergic neurons in the mammalian central nervous system *Physiol. Rev.*, 59 62, 1979

Björklund, A., and Stenevi, U Intracerebral neural implants Neuronal replacement and reconstruction of damaged circuitries *Annu Rev Neurosci.*, 7 279, 1984

Blumenthal, R., *et al.* (eds) Dynamic Patterns of Brain Cell Assemblies Neurosciences Research Program Bulletin Vol 12, No 1 Cambridge, Mass., Massachusetts Institute of Technology, 1974

Cotman, G. W., *et al* Synapse replacement in the nervous system of adult vertebrates. *Physiol Rev*, 61 684, 1981

Cowan, W M., The development of the brain *Sci Am* 241(3) 112, 1979

Darian, S. I. The Nervous System Section 1, Vol. III Sensory Processes. Baltimore, Waverly Press, 1984.

Freides, D., Human information processing and sensory modality Cross-modal functions, information complexity, memory, and deficit *Psychol. Bull.*, 81 284, 1974

Friesen, W O., and Stent, G S., Neural circuits for generating rhythmic movements. *Annu. Rev Biophys Bioeng*, 7 37, 1978

Gilbert, C. D.: Microcircuitry of the visual cortex *Annu Rev Neurosci*, 6 217, 1983.

Grinnell, A. D., Specificity of neurons and their interconnections. *In* Brookhart, J M, and Mountcastle, V B (eds) Handbook of Physiology Sec. 1, Vol 1 Baltimore, Williams & Wilkins, 1977 p 803

Henneman, E, and Mendell, L. M Functional organization of motoneuron pool and its inputs. *In* Brooks, V B (ed) Handbook of Physiology Sec. 1, Vol. II. Bethesda, American Physiological Society, 1981, p 423.

Kalia, M P., Anatomical organization of central respiratory neurons. *Annu Rev Physiol*, 43 105, 1981

Kandel, E R. Small systems of neurons *Sci Am* 241(3) 66, 1979

Karlin, A., *et al* (eds) Neuronal Information Transfer New York, Academic Press, 1978.

Kennedy, D., and Davis, W J Organization of invertebrate motor systems. *In* Brookhart, J M., and Mountcastle, V B (eds) Handbook of Physiology Sec. 1, Vol 1 Baltimore, Williams & Wilkins, 1977, p 1023

Laduron, P. M Presynaptic heteroreceptors in regulation of neuronal transmission. *Biochem Pharmacol*, 34 467, 1985

Levi Montalcini, R., and Calissano, P The nerve growth factor *Sci Am* 240(6) 68, 1979

Mendell, L M Modifiability of spinal synapses *Physiol. Rev.*, 64 260, 1984

Mountcastle, V B Central nervous mechanisms in mechanoreceptive sensibility *In* Darian-Smith, I (ed) Handbook of Physiology Sec. 1, Vol III Bethesda, American Physiological Society, 1984, p 789

Nicholls, J C., and Van Essen, D The nervous system of the leech *Sci Am*, 230 38, 1974

Pinsker, H M., and Willis, W D., Jr (eds) Information Processing in the *Nervous System. New York, Raven Press*, 1980

Purves, D., and Lichtman, J W Formation and maintenance of synaptic connections in autonomic ganglia. *Physiol Rev* 58 821, 1978

Purves, D and Lichtman, J W Specific connections between nerve cells *Annu Rev Physiol*, 45 553, 1983

Rovainen, C M Neurobiology of lampreys. *Physiol Rev.*, 59 1007, 1979.

Sachs, M B Neural coding of complex sounds Speech. *Annu Rev Physiol*, 46 261, 1984

Shepherd, B M The Synaptic Organization of the Brain. an Introduction New York, Oxford University Press, 1974

Shepherd, G M Microcircuits in the nervous system. *Sci. Am.*, 238(2) 92, 1978.

Sherman S M and Spear, P D Organization of visual pathways in normal and visually deprived cats *Physiol Rev.*, 62 738, 1982.

Sobel, L, *et al* Special techniques for the automatic computer reconstruction of neuronal structures. *Annu Rev Biophys Bioeng*, 9 347, 1980

Stein J F Introduction to Neurophysiology St Louis, C V Mosby, 1982

Sterling, P Microcircuitry of the cat retina *Annu Rev Neurosci.*, 6 149 1983

Su, C Purinergic neurotransmission and neuromodulation. *Annu Rev. Pharmacol Toxicol*, 23 297, 1983

Szentagothai, J, *et al.* (eds) *Conceptual Models of Neural Organization.* Neurosciences Research Program Cambridge, Mass, Massachusetts Institute of Technology, 1974

Uttley, A M. Information Transmission in the Nervous System New York. Academic Press, 1979

Wong, R. K., *et al* Local circuit interactions in synchronization of cortical neurones *J Exp Biol*, 112 169, 1984

Wooldridge, D E Sensory Processing in the Brain An Exercise in Neuroconnective Modeling New York, John Wiley & Sons, 1979

48

Sensory Receptors and Their Basic Mechanisms of Action

Input to the nervous system is provided by the sensory receptors that detect such sensory stimuli as touch, sound, light, pain, cold, warmth, and so forth. The purpose of this chapter is to discuss the basic mechanisms by which these receptors change sensory stimuli into nerve signals and, also, how both the type of sensory stimulus and its strength are detected by the brain.

TYPES OF SENSORY RECEPTORS AND THE SENSORY STIMULI THEY DETECT

The student of medical physiology will have already studied many different anatomical types of nerve endings, and Table 48–1 gives a list and classification of most of the body's sensory receptors. This table shows that there are basically five different types of sensory receptors: (1) *mechanoreceptors*, which detect mechanical deformation of the receptor or of cells adjacent to the receptor; (2) *thermoreceptors*, which detect changes in temperature, some receptors detecting cold and others warmth; (3) *nociceptors*, which detect damage in the tissues, whether it be physical damage or chemical damage; (4) *electromagnetic receptors*, which detect light on the retina of the eye; and (5) *chemoreceptors*, which detect taste in the mouth, smell in the nose, oxygen level in the arterial blood, osmolality of the body fluids, carbon dioxide concentration, and perhaps other factors that make up the chemistry of the body.

This chapter will discuss the function of a few specific types of receptors, primarily peripheral mechanoreceptors, to illustrate some of the basic principles by which receptors in general operate. Other receptors will be discussed in relation to the sensory systems that they subserve. Figure 48–1 illustrates some of the different types of mechanoreceptors found in the skin or in the deep structures of the body, and Table 48–1 gives their respective sensory functions. All these receptors will be discussed in the following chapters in relation to the respective sensory systems. However, the functions of some of these are described for illustrative purposes in this chapter as well.

DIFFERENTIAL SENSITIVITY OF RECEPTORS

The first question that must be answered is, how do two types of sensory receptors detect different types of sensory stimuli? The answer is by virtue of differential sensitivities. That is, each type of receptor is very highly sensitive to one type of

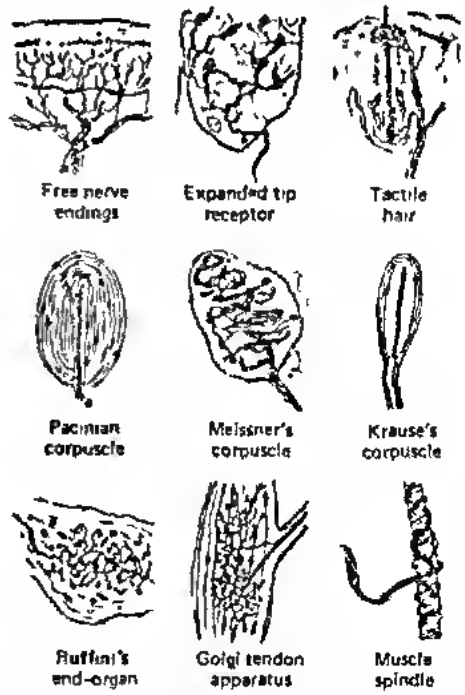


Figure 48–1. Several types of somatic sensory nerve endings

Table 48–1 CLASSIFICATION OF SENSORY RECEPTORS

Mechanoreceptors
Skin tactile sensibilities (epidermis and dermis)
- Free nerve endings
- Expanded tip endings
  - Merkel's discs
  - Plus several other variants
- Spray endings
  - Ruffini's endings
- Encapsulated endings
  - Meissner's corpuscles
  - Krause's corpuscles
- Hair end-organs

Deep tissue sensibilities
- Free nerve endings
- Expanded tip endings
  - Plus a few other variants
- Spray endings
  - Ruffini's endings
- Encapsulated endings
  - Pacinian corpuscles
  - Plus a few other variants
- Muscle endings
  - Muscle spindles
  - Golgi tendon receptors

Hearing
- Sound receptors of cochlea

Equilibrium
- Vestibular receptors

Arterial pressure
- Baroreceptors of carotid sinuses and aorta

Thermoreceptors
Cold
- Cold receptors

Warmth
- Warm receptors

Nociceptors
Pain
- Free nerve endings

Electromagnetic Receptors
Vision
- Rods
- Cones

Chemoreceptors
Taste
- Receptors of taste buds

Smell
- Receptors of olfactory epithelium

Arterial oxygen
- Receptors of aortic and carotid bodies

Osmolality
- Probably neurons of supraoptic nuclei

Blood CO_2
- Receptors in or on surface of medulla and in aortic and carotid bodies

Blood glucose, amino acids, fatty acids
- Receptors in hypothalamus

stimulus for which it is designed and yet is almost nonresponsive to normal intensities of the other types of sensory stimuli. Thus, the rods and cones are highly responsive to light but are almost completely nonresponsive to heat, cold, pressure on the eyeballs, or chemical changes in the blood. The osmoreceptors of the supraoptic nuclei in the hypothalamus detect minute changes in the osmolality of the body fluids but have never been known to respond to sound. Finally, pain receptors in the skin are almost never stimulated by usual touch or pressure stimuli but do become highly active the moment tactile stimuli become severe enough to damage the tissues.

Modality of Sensation—The "Labeled Line" Principle

Each of the principal types of sensation that we can experience—pain, touch, sight, sound, and so forth—is called a *modality* of sensation. Yet, despite the fact that we experience these different modalities of sensation, nerve fibers transmit only impulses Therefore, how is it that different nerve fibers transmit different modalities of sensation?

The answer to this is that each nerve tract terminates at a specific point in the central nervous system, and the type of sensation felt when a nerve fiber is stimulated is determined by this point in the nervous system to which the fiber leads. For instance, if a pain fiber is stimulated, the person perceives pain regardless of what type of stimulus excites the fiber. The stimulus can be electricity, heat, crushing, or stimulation of the pain nerve ending by damage to the tissue cells. Yet, whatever the means of stimulation, the person still perceives pain. Likewise, if a touch fiber is stimulated by exciting a touch receptor electrically or in any other way, the person perceives touch because touch fibers lead to specific touch areas in the brain. Similarly, fibers from the retine of the eye terminate in the vision areas of the brain, fibers from the ear terminate in the auditory areas of the broin, and temperature fibers terminate in the temperature areas

This specificity of nerve fibers for transmitting only one modality of sensation is called the *"labeled line" principle.*

TRANSDUCTION OF SENSORY STIMULI INTO NERVE IMPULSES

LOCAL CURRENTS AT NERVE ENDINGS—RECEPTOR POTENTIALS

All sensory receptors have one feature in common. Whatever the type of stimulus that excites the receptor, its immediate effect is to change the potential across the receptor membrane. This change in potential is called a *receptor potential.*

Mechanisms of Receptor Potentials. Different receptors can be excited in several different ways to cause receptor potentials: (1) by mechanical deformation of the receptor, which stretches the membrane and opens ion channels; (2) by application of a chemical to the membrane, which also opens ion channels; (3) by change of the temperature of the membrane, which alters the permeability of the membrane; or (4) by the effects

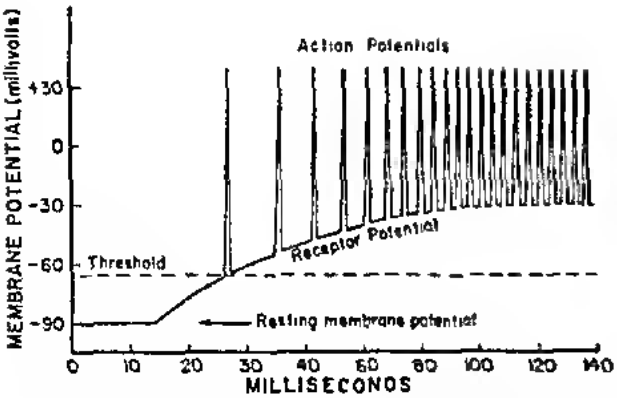


Figure 48–2. Typical relationship between receptor potential and action potentials when the receptor potential rises above the threshold level

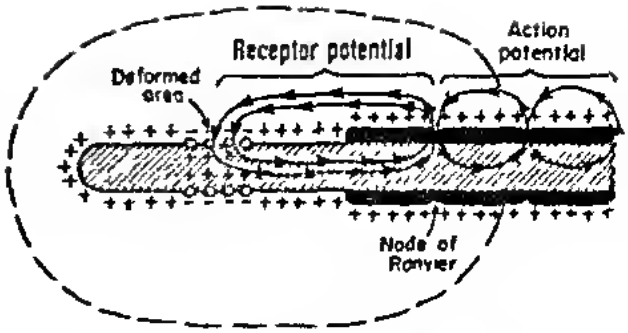


Figure 48–3. Excitation of a sensory nerve fiber by a receptor potential produced in a pacinian corpuscle (Modified from Loewenstein: *Ann NY Acad Sci*, 94 510, 1961)

of electromagnetic radiation such as light on the receptor, which either directly or indirectly changes the membrane characteristics and allows ions to flow through membrane channels. It will ba recognized that these four different means for exciting receptors correspond in general with the different types of known sensory receptors In all instances, the basic cause of the change in membrana potential is a change in receptor membrane permeability, which allows ions to diffuse more or less readily through the membrane and thereby change the transmembrane potential.

Relationship of the Receptor Potential to Action Potentials. When the receptor potential rises above the threshold for eliciting action potentials in the nerve fiber attached to the receptor, then action potentials begin to appear. This is illustrated in Figure 48–2. Note also that the more the receptor potential rises above the threshold level, the greater becomes the action potential frequency Thus, the receptor potential stimulates the terminal nerve fiber in the same way that the excitatory postsynaptic potential stimulates neurons in the central nervous system.

Illustrative Examples of Receptor Function

The Receptor Potential of the Pacinian Corpuscle. The pacinian corpuscle is a very large and easily dissected sensory receptor. For this reason, one can study in detail the mechanism by which tactile stimuli excite it and by which it causes action potentials in the sensory nerve fiber leading from it. Note in Figure 48–3 that the pacinian corpuscle has a central nonmyelinated tip of a nerve fiber extending through its core. Surrounding this fiber are many concentric capsule layers, as shown in Figure 48–1, so that compression on the outside of the corpuscle tends to elongate, indent, or otherwise deform the central core of the fiber, depending on how the compression is applied. This deformation causes sudden opening of ion channels that mainly carry positively charged sodium ions to the interior of the fiber. This in turn creates more positivity in the fiber, which is the receptor potential. The receptor potential in turn induces a local circuit of current flow that spreads along the nerve fiber At the first node of Ranvier, which itself lies inside the capsule of the pacinian corpuscle, the local current flow initiates action potentials in the fiber That is, the current flow through the node depolarizes it, and this then sets off a typical action potential that is transmitted along the nerve fiber toward the central nervous aystem

Relatianshlp Between Stimulus Strength and the Receptor Potenttial. Figure 48–4 illustrates the effect on the amplitude of the receptor potential caused by progressively stronger mechanical compression applied experimentally to the central core of the pacinian corpuscle. Note that the amplitude increases rapidly at first but then progressively less rapidly at high stimulus strengths. The maximum amplitude that can be achieved by receptor potentials is around 100 millivolts. That is, a receptor potential can have almost as high a voltage as an action potential.

Relationship of Amplitude of Receptor Potential to Nerve Impulse Rnte. The frequenry of

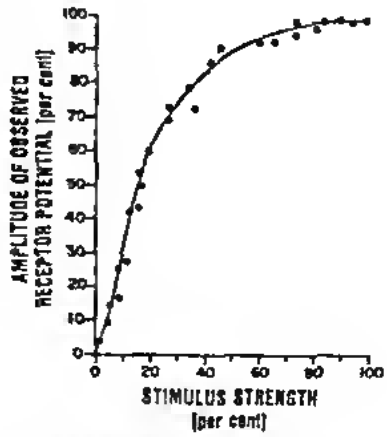


Figure 48–4. Relationship of amplitude of receptor potential to strength of a stimulus applied to a pacinian corpuscle. (From Loewenstein *Ann. N.Y. Acad Sci*, 94 510, 1961)

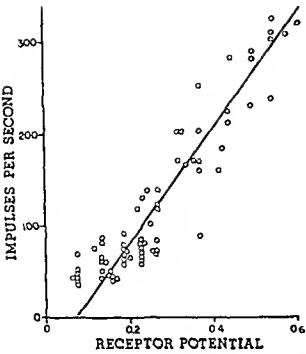


Figure 48–5. Relationship between the receptor potential of a muscle spindle and the frequency of sensory impulses transmitted from the spindle. (from Katz, *J Physiol. (Lond.), 111.261, 1950*)

action potentials in the nerve fiber (impulse rate) is almost directly proportional to the amplitude of the receptor potential. This relationship is illustrated in Figure 48–5, which shows the impulse rate corresponding to different receptor potential amplitudes recorded from a muscle spindle; there is an almost exact proportional relationship. This same relationship between receptor potential and impulse rate is approximately true for most sensory receptors.

ADAPTATION OF RECEPTORS

A special characteristic of all sensory receptors is that they *adapt* either partially or completely to their stimuli after a period of time. That is, when a continuous sensory stimulus is applied, the receptors respond at first with a very high impulse rate, then at a progressively lower rate until finally many of them no longer respond at all.

Figure 48–6 illustrates typical adaptation of certain types of receptors. Note that the pacinian corpuscle adapts extremely rapidly and hair receptors adapt within a second or so, while joint capsule and muscle spindle receptors adapt very slowly.

Furthermore some sensory receptors adapt to a far greater extent than others. For example, the pacinian corpuscles adapt to "extinction" within a few thousandths to a few hundredths of a second, and the hair base receptors adapt to extinction within a second or more. It is probable that all other *mechanoreceptors* also adapt completely eventually, but some require hours or days to do so, for which reason they are frequently called "nonadapting" receptors. The longest measured time for complete adaptation of a mechanoreceptor is about two days for the carotid and aortic baroreceptors.

Some of the other types of receptors, the chemoreceptors and pain receptors for instance, probably never adapt completely. Each of these will be discussed in turn in subsequent chapters.

Mechanisms by Which Receptors Adapt. Adaptation of receptors is an individual property of each type of receptor in much the same way that development of a receptor potential is an indivdual property. For instance, in tha eye, the rods and cones adapt by changing their chemical compositions (which will be discussed in Chapter 59). In the case of the mechanoreceptors, the receptor that has been studied for adaptation in greatest detail is the pacinian corpuscle. Adaptation occurs in this receptor in two ways. First, the corpuscular structure itself very rapidly adapts to the *deformation* of the tissue. This can be explained as follows: The pacinian corpuscle is a viscoelastic structure so that when a distorting force is suddenly applied to one side of tha corpuscla this force is instantly transmitted by the viscous component of tha corpuscle directly to tha same side of the central core, thus eliciting the *receptor potential*. However, within a few thousandths to a few hundredths of a second the fluid within tha corpuscle redistributes so that the pressure becomes essentially equal all through the corpuscle; this now applies an even pressure on oll sides of the central core fiber, so that the receptor potential is no longer elicited. Thus, the *receptor potential* appears at the onset of compression but then disappears within a small fraction of a second even though the compression continues.

Then, when the distorting force is removed from tho corpuscle, essentially the reverse events occur. The sudden removal of the distortion from one side of the corpuscle allows rapid expansion on that side, and a corresponding distortion of the central coro occurs once more. Again, within milliseconds,

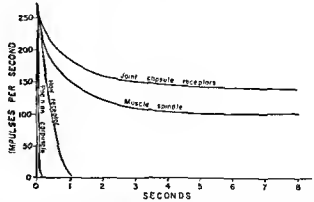


Figure 48–6. Adaptation of different types of receptors, showing rapid adaptation of some receptors and slow adaptation of others.

the pressure becomes equalized all through the corpuscle and the stimulus is lost Nevertheless, this disturbance of the central core fiber signals the offset of compression as well as signalling the onset of compression.

The second mechanism of adaptation of the pacinian corpuscle, but a much slower one, results from a process called *accommodation* that occurs in the nerve fiber itself. That is, even if by chance the central core fiber should continue to be excited, as can be achieved after the capsule has been removed and the core is compressed with a stylus, the tip of the nerve fiber itself gradually becomes "accommodated" to the stimulus This probably results from "inactivation" of the sodium channels in the nerve fiber membrane, which means that the current flow itself through the channels in some way causes them gradually to close, as was explained in Chapter 10

Presumably, these same two general mechanisms of adaptation apply to other types of mechanoreceptors. That is, part of the adaptation results from readjustments in the structure of the receptor itself, and part results from accommodation in the terminal nerve fibril.

Function of the Slowly Adapting and Nonadapting Receptors to Detect Continuous Stimulus Strength—The "Tonic" Receptors. The slowly adapting receptors continue to transmit impulses to the brain as long as the stimulus is present (or at least for many minutes or hours). Therefore, they keep the brain constantly apprised of the status of the body and its relation to its surroundings. For instance, impulses from the slowly adapting joint capsule receptors allow the person to "know" at all times the degree of bending of the joints and therefore the positions of the different parts of the body. And impulses from the muscle spindles and Golgi tendon apparatuses allow the central nervous system to know respectively the status of muscle contraction and the load on the muscle tendon at each instant

Other types of slowly adapting receptors include the receptors of the macula in the vestibular apparatus, the pain receptors, the baroreceptors of the arterial tree, the chemoreceptors of the carotid and aortic bodies, and some of the tactile receptors, such as the Ruffini endings and the Merkel's discs.

Because the slowly adapting receptors can continue to transmit information for many hours, they are also called *tonic* receptors. Many of these slowly adapting receptors will adapt to extinction if the intensity of the stimulus remains absolutely constant for several hours or days Fortunately, because of our continually changing bodily state, these receptors almost never reach a state of complete adaptation

Function of the Rapidly Adapting Receptors to Detect Change in Stimulus Strength—The "Rate Receptors" or "Movement Receptors" or "Phasic Receptors." Obviously, receptors that adapt rapidly cannot be used to transmit a continuous signal because these receptors are stimulated only when the stimulus strength changes. Yet they react strongly *while a change is actually taking place.* Furthermore, the number of impulses transmitted is directly related to the *rate at which the change takes place.* Therefore, these receptors are called *rate* receptors, *movement* receptors, or *phasic* receptors. Thus, in the case of the pacinian corpuscle, sudden pressure applied to the skin excites this receptor for a few milliseconds, and then its excitation is over even though the pressure continues But then it transmits a signal again when the pressure is released. In other words, the pacinian corpuscle is exceedingly important for transmitting information about rapid changes in pressure against the body, but it is useless for transmitting information about constant pressure applied to the body.

Importance of the Rate Receptors—Their Predictive Function. If one knows the rate at which some change in bodily status is taking place, one can predict the state of the body a few seconds or even a few minutes later For instance, the receptors of the semicircular canals in the vestibular apparatus of the ear detect the rate at which the head begins to turn when one runs around a curve. Using this information, a person can predict a turn of 10, 30, or some other number of degrees within the next 10 seconds and can adjust the motion of the limbs *ahead of time* to keep from losing balance. Likewise, pacinian corpuscles and other receptors located in or near the joint capsules help detect the rates of movement of the different parts of the body Therefore, when one is running, information from these receptors allows the nervous system to predict where the feet will be during any precise fraction of a second, and appropriate motor signals can be transmitted to the muscles of the legs to make any necessary anticipatory corrections in limb position so that the person will not fall. Loss of this predictive function makes it impossible for the person to run

PSYCHIC INTERPRETATION OF STIMULUS STRENGTH

The ultimate goal of most sensory stimulation is to apprise the psyche of the state of the body and its surroundings. Therefore, it is important that we discuss briefly some of the principles related to the transmission of sensory stimulus strength to the higher levels of the nervous system

The first question that comes to mind is. How is it possible for the sensory system to transmit sensory experiences of tremendously varying intensities? For instance, the auditory system can detect the weakest possible whisper but can also discern the meanings of an explosive sound only a few feet away, even though the sound intensities of these two experiences can vary more than ten billion fold, the eyes can see visual images with light intensities that vary as much as a half million

fold; or the skin can detect pressure differences of ten thousand- to one hundred thousandfold.

As a partial explanation of these effects, note in Figure 48–4 the relationship of the receptor potential produced by the pacinian corpuscle to the strength of stimulus. *At low stimulus strength,* very slight changes in stimulus strength increase the potential markedly, whereas at high levels of stimulus strength further increases in receptor potential are very slight. Thus, the pacinian corpuscle is capable of accurately measuring extremely minute changes in stimulus strength at low intensity levels, but at high intensity levels the change in stimulus strength must be much greater to cause the same amount of change in receptor potential.

The transduction mechanism for detecting sound by the cochlea of the ear illustrates still another method for separating gradations of stimulus intensity. When sound causes vibration at a specific point on the basilar membrane, weak vibration stimulates only those hair cells at the point of maximum vibration. But, as the vibration intensity increases, not only do these hair cells become more intensely stimulated, but still many more hair cells in each direction farther away from the maximum vibratory point also become stimulated Thus, signals transmitted over progressively increasing numbers of cochlear nerve fibers is another mechanism by which stimulus strength is transmitted into the central nervous system This mechanism plus the direct effect of stimulus strength on impulse rate in each nerve fiber, as well as several other mechanisms, makes it possible for the ear to operate reasonably faithfully at stimulus intensity levels changing more than ten billionfold

Importance of the Tremendous Intensity Range of Sensory Reception. Were it not for this tremendous intensity range of sensory reception that we can experience, the various sensory systems would more often than not operate in the wrong range This is illustrated by the attempts of most persons to adjust the light exposure on a camera without using a light meter Left to intuitive judgment of light intensity, a person almost always overexposes the film on very bright days and greatly underexposes the film at twilight Yet, that person's eyes are capable of discriminating with great detail the surrounding objects in both very bright sunlight and at twilight, the camera cannot do this because of the narrow critical range of light intensity required for proper exposure of film.

JUDGMENT OF STIMULUS STRENGTH

Physiopsychologists have evolved numerous methods for testing one's judgment of sensory stimulus strength, but only rarely do the results from the different methods agree with each other For instance, one testing method requests the subject to select a weight that is exactly 100 per cent heavier than another But the person usually selects a weight that is about 50 per cent heavier instead of 100 per cent heavier. Thus, the weak stimulus is underestimated and the strong stimulus is overestimated. In still another test procedure, a person is given a weight to hold and is then required to select the minimum amount of additional weight that must be added to detect a difference In this case the person might be holding a 30 gram weight and finds that an additional 1 gram is needed to detect a difference. Then the person holds a 300 gram weight and finds that 10 grams of additional weight are required In this instance the discriminatory ability is far greater at the low intensity level than at the high intensity level. Thus, the results of these two different types of tests are exactly opposite to each other, which means that the real argument lies in the meaning of the tests themselves. Therefore, at present no real agreement exists as to a proper method for measuring one's judgment of stimulus strength. Yet, two principles are widely discussed in the physiopsychology field of sensory interpretation: the *Weber-Fechner principle* and the *power principle.*

The Weber-Fechner Principle—Detection of "Ratio" of Stimulus Strength. In the mid-1800s, Weber first and Fechner later proposed the principle that *gradations of stimulus strength are discriminated approximately in proportion to the logarithm of stimulus strength.* This principle is based primarily on one's ability to judge minimal *changes* in stimulus strength. That is, in the second test described in the previous section the person could barely detect a 1 gram increase in weight when holding 30 grams, or a 10 gram increase when holding 300 grams. Thus, the *ratio* of the change in stimulus strength required for detection of a change remained essentially constant, about 1 to 30, which is what the logarithmic principle means. To express this mathematically,

$$\text{Interpreted signal strength} = \log(\text{Stimulus}) + \text{Constant}$$

Because tha Weber-Fechner principla offers a ready explanation for the tremendous range of stimulus strength that our nervous system can discern, it unfortunately became widely accepted for all types of sensory experience and for all levels of background sensory intensity More recently it has become evident that this principle applies mainly to higher intensities of visual, auditory, and cutaneous sensory experience and it applies only poorly to most other types of sensory experience.

Yet, the Weber-Fechner principla is still a good ona to remember because it emphasizes that the greater the background sensory stimulus, the greater also must ba the additional change in stimulus strength in order for the psyche to detect the change

The Power Law. Another attempt by physiopsychologists to find a good mathematical relationship between actual stimulus strength and interpretation of stimulus strength is the following formula, known as the power law

$$\text{Interpreted signal strength} = K \cdot (\text{Stimulus} - k)^y$$

In this formula K and k are constants, and y is the power to which the stimulus strength is raised. The exponent y and the constants K and k are different for each type of sensation.

When this power law relationship is plotted on a graph using double logarithmic coordinates, as illustrated in Figure 48–7, a linear relationship can be attained between interpreted stimulus strength and actual stimulus strength over a large range for almost any type of sensory perception. However, as illustrated in the figure, even this power law relationship fails to hold satisfactorily at both very low and very high stimulus strengths.

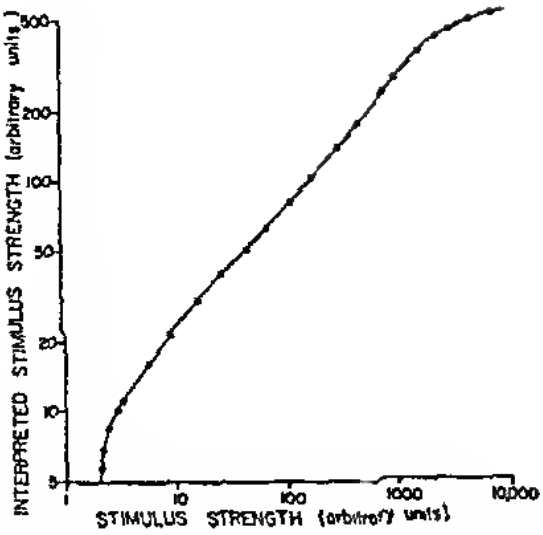


Figure 48–7. Graphical demonstration of the "power law" relationship between actual stimulus strength and strength that the psyche interprets it to be. Note that the power law does not hold at either very weak or very strong stimulus strengths.

PHYSIOLOGICAL CLASSIFICATION OF NERVE FIBERS

Some sensory signals need to be transmitted to the central nervous system extremely rapidly, or otherwise the information would be useless. An example of this is the sensory signals that apprise the brain of the momentary positions of the limbs at each fraction of a second during running. It is important that this information be transmitted by way of extremely rapidly conducting nerve fibers. At the other extreme, some types of sensory information, such as that depicting prolonged, aching pain, do not need to be transmitted rapidly at all, so that very slowly conducting fibers will suffice. Fortunately, nerve fibers come in all sizes between 0.2 and 20 microns in diameter—the larger the diameter, the greater the conducting velocity. The range of conducting velocities is between 0.5 and 120 meters per second.

Figure 48–8 gives two different classifications of nerve fibers that are in general use. One of these is a general classification that includes both sensory and motor fibers, including the autonomic nerve fibers as well. The other is a classification of sensory nerve fibers that is used primarily by sensory neurophysiologists.

In the general classification, the fibers are divided into types A and C, and the type A fibers are further subdivided into α, β, γ, and δ fibers.

Type A fibers are the typical myelinated fibers of spinal nerves. Type C fibers are the very small, unmyelinated nerve fibers that conduct impulses at low velocities. These constitute more than half the sensory fibers in most peripheral nerves and also all of the postganglionic autonomic fibers.

The sizes, velocities of conduction, and functions of the different nerve fiber types are given in the figure. Note that the very large fibers can transmit impulses at velocities as great as 120 meters per second, a distance in one second that is longer than a football field. On the other hand, the smallest fibers transmit impulses as slowly as 0.5 meter per second, requiring about 2 seconds to go from the big toe to the spinal cord.

Over two thirds of all the nerve fibers in peripheral nerves are type C fibers. Because of their great number, these can transmit tremendous amounts of information from the surface of the body, even though their velocities of transmission are very slow. Utilization of type C fibers for transmitting this great mass of information represents an important economy of space in the nerves, for use of the larger type A fibers for transmitting all information would require peripheral nerves the size of large ropes and a spinal cord almost as large as the body itself.

Alternate Classification Used by Sensory Physiologists. Certain recording techniques have made it possible to separate the type Aα fibers into two subgroups; and, yet, these same recording techniques cannot distinguish easily between Aβ and Aγ fibers. Therefore, the following classification is frequently used by sensory physiologists:

Group Ia. Fibers from the annulospiral endings of muscle spindles (Average about 17 microns in diameter. These are alpha type A fibers in the general classification.)

Group Ib. Fibers from the Golgi tendon organs (Average about 16 microns in diameter; these also are alpha type A fibers.)

Group II. Fibers from the discrete cutaneous tactile

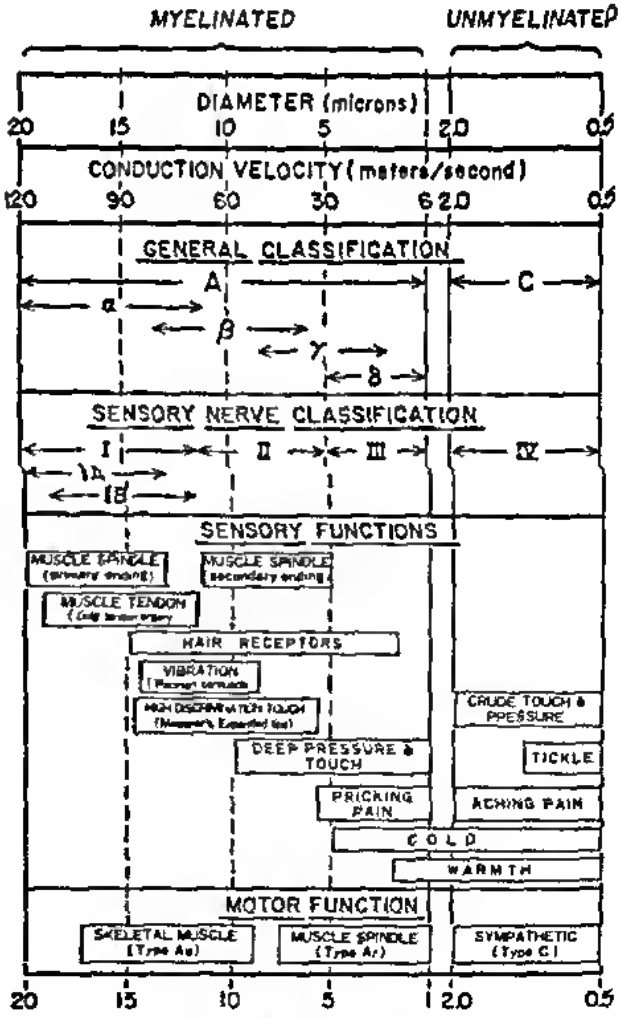


Figure 48–8. Physiological classifications and functions of nerve fibers.

receptors and also from the flower-spray endings of the muscle spindles. (Average about 8 microns in diameter; these are beta and gamma type A fibers in the other classification.)

Group III. Fibers carrying temperature, crude touch, and pricking pain sensations. (Average about 3 microns in diameter; these are delta type A fibers in the other classification.)

Group IV. Unmyelinated fibers carrying pain, itch, temperature, and crude touch sensations. (0.5 to 2 microns in diameter; called type C fibers in the other classification.)

REFERENCES

American Physiological Society Sensory Processes Washington, D C, American Physiological Society, 1984

Anderson, D J, *et al* Sensory mechanisms in mammalian teeth and their supporting structures. *Physiol Rev*, *50*:171, 1970

Anstis, S M, *et al* Perception New York, Springer-Verlag, 1978

Babel, J, *et al*. Ultrastructure of the Peripheral Nervous System and Sense Organs New York, Churchill Livingstone, 1971

Bate, C M, *et al*: Development of Sensory Systems New York, Springer-Verlag, 1978

Bennett, T. L The Sensory World An Introduction to Sensation and Perception Monterey, Cal, Brooks/Cole Publishing Co, 1978

Brown, E, and Deffenbacher, K Perception and the Senses New York, Oxford University Press, 1979

Brown, M Touch Will Tell New York, Franklin Watts, 1979

Catton, W T Mechanoreceptor function *Physiol Rev*, *50*:297, 1970

Coren, S., *et al* Sensation and Perception New York, Academic Press, 1979

Darian-Smith, I The sense of touch Performance and peripheral neural processes *In* Darian-Smith, I (ed) Handbook of Physiology Sec 1, Vol. III Bethesda, American Physiological Society, 1984, p 739

Freedhoff, A J, and Miller, J C Clinical implications of receptor sensitivity modification *Annu Rev Neurosci*, *6* 121, 1983

Goldstein, E B Sensation and Perception Belmont, Cal, Wadsworth Publishing Co, 1980

Grant, R Receptors and Sensory Perception New Haven, Conn, Yale University Press, 1955

Gray, J. A B: Initiation of impulses at receptors. *In* Magoun, H W (ed) Handbook of Physiology Sec. 1, Vol I Baltimore, Williams & Wilkins, 1959

Halata, Z: The Mechanoreceptors of the Mammalian Skin New York, Springer-Verlag, 1975

Hochberg, J. Perception. *In* Darian-Smith, I (ed) Handbook of Physiology Sec 1, Vol III Bethesda, American Physiological Society, 1984, p 75

Jung, R. Sensory research in historical perspective Some philosophical foundations of perception *In* Darian-Smith, I. (ed.) Handbook of Physiology. Sec. 1, Vol III Bethesda, American Physiological Society, 1984, p 1

Loewenstein, W R. The generation of electrical activity in a nerve ending *Ann N Y Acad Sci*, *81*:367, 1959

Loewenstein, W R. Excitation and inactivation in a receptor membrane *Ann N Y Acad. Sci.*, *94* 510, 1961

Lynn, B Somatosensory receptors and their CNS connections *Annu Rev Physiol*, *37* 105, 1975

McCloskey, D I Kinesthetic sensibility *Physiol Rev*, *58* 763, 1978

Neff, W D. (ed) Contributions to Sensory Physiology New York, Academic Press, 1982

Pantal, A A. (ed) Morphology and Mechanisms of Chemoreceptors. Delhi, India, Vallabhbhai Patel Chest Institute, University of Delhi, 1976

Porter, R. (ed) Studies in Neurophysiology New York, Cambridge University Press, 1978

Schmidt, R F (ed) Fundamentals of Sensory Physiology New York, Springer Verlag, 1978

Somjen, G Sensory Coding in the Mammalian Nervous System. New York, Appleton Century-Crofts, 1972

Stebbins W C Brown, C H, and Petersen, M R. Sensory function in animals *In* Darian Smith, I (ed) Handbook of Physiology Sec. 1, Vol III Bethesda, American Physiological Society, 1984, p 123.

Thompson, R F, and Patterson, M (eds.) Bioelectric Recording Techniques Part C Receptor and Effector Processes. New York, Academic Press, 1974

Weiss, T F Relation of receptor potentials of cochlear hair cells to spike discharges of cochlear neurons *Annu Rev Physiol*, *46* 247, 1984

Wiederhold, M L Mechanosensory transduction in "sensory" and "motile" cilia. *Annu Rev Biophys Bioeng*, *5* 39, 1976

Wiersma, C A G, and Roach, J L. M Principles in the organization of invertebrate sensory systems. *In* Brookhart, J M, and Mountcastle, V B (eds) *Handbook of Physiology* Sec 1, Vol 1 Baltimore, Williams & Wilkins, 1977, p 1089

Zimmermann, M Neurophysiology of nociception *Int Rev Physiol.*, *10* 179, 1978

49

Somatic Sensations: I. The Mechanoreceptive Sensations

The *somatic senses* are the nervous mechanisms that collect sensory information from the body. These senses are in contradistinction to the *special senses*, which mean specifically vision, hearing, smell, taste, and equilibrium.

CLASSIFICATION OF SOMATIC SENSES

The somatic senses can be classified into three different physiological types: (1) the *mechanoreceptive somatic senses*, stimulated by mechanical displacement of some tissue of the body, (2) the *thermoreceptive senses*, which detect heat and cold, and (3) the *pain sense*, which is activated by any factor that damages the tissues. This chapter deals with the mechanoreceptive somatic senses, and the following chapter discusses the thermoreceptive and pain senses.

The mechanoreceptive senses include *touch*, *pressure*, *vibration*, and *tickle* senses (which are frequently called the *tactile senses*) and the *position sense*, which determines the relative positions and rates of movement of the different parts of the body.

Other Classifications of Somatic Sensations. Different types of somatic sensations are also grouped together in special classes that are not necessarily mutually exclusive, as follows:

Exteroceptive sensations are those from the surface of the body. *Proprioceptive sensations* are those having to do with the physical state of the body, including position sensations, tendon and muscle sensations, pressure sensations from the bottom of the feet, and even the sensation of equilibrium, which is generally considered to be a "special" sensation rather than a somatic sensation.

Visceral sensations are those from the viscera of the body; in using this term one usually refers specifically to sensations from the internal organs.

The *deep sensations* are those that come from the deep tissues, such as from fasciae, muscles, bone, and so forth. These include mainly "deep" pressure, pain, and vibration.

DETECTION AND TRANSMISSION OF TACTILE SENSATIONS

Interrelationship Between the Tactile Sensations of Touch, Pressure, and Vibration. Though touch, pressure, and vibration are frequently classified as separate sensations, they are all detected by the same types of receptors. The only differences among these three are (1) touch sensation generally results from stimulation of tactile receptors in the skin or in tissues immediately beneath the skin, (2) pressure sensation generally results from deformation of deeper tissues, and (3) vibration sensation results from rapidly repetitive sensory signals, but some of the same types of receptors as those for touch and pressure are utilized—specifically the very rapidly adapting types of receptors.

The Tactile Receptors. At least six entirely different types of tactile receptors are known, but many more similar to these also exist. Some of these receptors were illustrated in Figure 48–1 of Chapter 48, and their special characteristics are the following:

First, some *free nerve endings*, which are found everywhere in the skin and in many other tissues, can detect touch and pressure. For instance, even light contact with the cornea of the eye, which contains no other type of nerve ending besides free nerve endings, can nevertheless elicit touch and pressure sensations.

Second, a touch receptor of special sensitivity is *Meissner's corpuscle*, an encapsulated nerve ending that excites a large (type Aβ) myelinated sensory nerve fiber. Inside the capsulation are many whorls of terminal nerve filaments. These receptors are present in the nonhairy parts of the skin (called *glabrous skin*) and are particularly abundant in the fingertips, lips, and other areas of the skin where one's ability to discern spatial characteristics of touch sensations is highly developed.

They, along with the expanded tip receptors described subsequently, are mainly responsible for the ability to recognize exactly what point of the body is touched and to recognize the texture of objects touched. Meissner's corpuscles adapt in a fraction of a second after they are stimulated, which means that they are particularly sensitive to movement of very light objects over the surface of the skin and also to low frequency vibration.

Third, the fingertips and other areas that contain large numbers of Meissner's corpuscles also contain large numbers of *expanded tip tactile receptors*, one type of which is *Merkel's discs*. The hairy parts of the skin also contain moderate numbers of expanded tip receptors, even though they have almost no Meissner's corpuscles. These receptors differ from Meissner's corpuscles in that they transmit an initially strong but partially adapting signal and then a continuing weaker signal that adapts only slowly. Therefore, they are responsible for giving steady state signals that allow one to determine continuous touch of objects against the skin. Merkel's discs are often grouped together in a single receptor organ called the *Iggo dome receptor*, which projects upward against the undersids of the epithelium of the skin, as illustrated in Figure 49–1. This causes the epithelium st this point to protrude outward, thus creating a dome. Also note that the entire group of Merkel's discs is innervated by a single large type of myelinated nerve fiber (type Aβ). Each dome receptor has a diameter averaging only 0.2 mm, and these receptors are extremely sensitive. They probably play an important role in localizing touch sensations to the specific surface areas of the body.

Fourth, slight movement of any hair on the body stimulates the nerve fiber entwining its base. Thus, each hair and its basal nerve fiber, called the *hair end-organ*, is also a touch receptor. This receptor adapts readily and, therefore, like Meissner's corpuscles, detects mainly movement of objects on the surface of the body or initial contact with the body.

Fifth, located in the deeper layers of the skin and also in deeper tissues are many *Ruffini's end-organs*, which are multibranched, encapsulated endings, as described and illustrated in the previous chapter. These endings adapt very little and, therefore, are important for signaling continuous states of deformation of the skin and deeper tissues, such as heavy and continuous touch signals and pressure signals. They are also found in joint capsules and help signal the degree of joint rotation.

Sixth, *pacinian corpuscles*, which were discussed in detail in Chapter 48, lie both immediately beneath the skin and also deep in the fascial tissues of the body. These are stimulated only by very rapid movement of the tissues because they adapt in a few thousandths or a few hundredths of a second. Therefore, they are particularly important for detecting tissue vibration or other extremely rapid changes in the mechanical state of the tissues.

Transmission of Tactile Sensations in Peripheral Nerve Fibers. Almost all the specialized sensory receptors, such as Meissner's corpuscles, Iggo dome receptors, hair receptors, pacinian corpuscles, and Ruffini's endings, transmit their signals in type Aβ nerve fibers that have transmission velocities of 30 to 70 meters per second On the other hand, free nerve ending tactile receptors transmit signals mainly via the small type Aδ myelinated fibers that conduct at velocities of 5 to 30 meters per second. Some tactile free nerve endings transmit via type C unmyelinated fibers at velocities from a fraction of a meter up to 2 meters per second; these send signals into the spinal cord and lower brain stem, probably subserving mainly the sensation of tickle. Thus, the

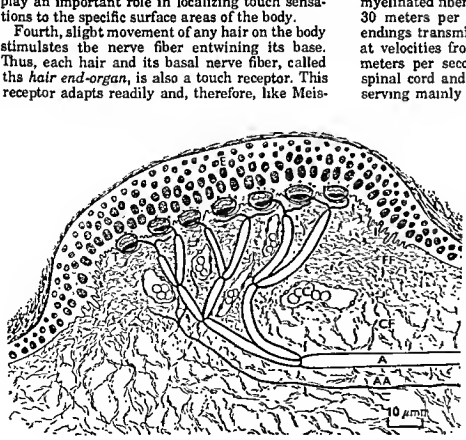


Figure 49–1. The Iggo dome receptor. Note the multiple numbers of Merkel's discs innervated by a single large myelinated fiber and abutting tightly the undersurface of the epithelium. (From Iggo and Muir, *J Physiol.*, 200:763, 1969.)

more critical types of sensory signals—those that help to determine precise localization on the skin, minute gradations of intensity, or rapid changes in sensory signal intensity—are all transmitted in the rapidly conducting types of sensory nerve fibers. On the other hand, the cruder types of signals, such as crude pressure, poorly localized touch, and especially tickle, are transmitted via much slower nerve fibers that require much less space than the faster fibers.

DETECTION OF VIBRATION

All the different tactile receptors are involved in detection of vibration, though different receptors detect different frequencies of vibration. Pacinian corpuscles can signal vibrations from 60 to 500 cycles per second, because they respond extremely rapidly to minute and rapid deformations of the tissues, and they also transmit their signals over type Aβ nerve fibers, which can transmit more than 1000 impulses per second.

Low frequency vibrations up to 80 cycles per second, on the other hand, stimulate other tactile receptors—especially Meissner's corpuscles, which are less rapidly adapting than pacinian corpuscles.

TICKLING AND ITCH

Recent neurophysiological studies have demonstrated the existence of very sensitive, rapidly adapting, mechanoreceptive free nerve endings that elicit only the tickle and itch sensation. Furthermore, these endings are found almost exclusively in the superficial layers of the skin, which is also the only tissue from which the tickle and itch sensation usually can be elicited. This sensation is transmitted by very small type C, unmyelinated fibers similar to those that transmit the aching type of pain.

The purpose of the itch sensation is presumably to call attention to mild surface stimuli such as a flea crawling on the skin or a fly about to bite, and the elicited signals then excite the scratch reflex or other maneuvers that rid the host of the irritant.

The relief of itch by the process of scratching occurs only when the irritant is removed or when the scratch is strong enough to elicit pain. The pain signals are believed to suppress the itch signals in the cord by the process of inhibition that will be described in the following chapter.

THE DUAL SYSTEM FOR TRANSMISSION OF MECHANORECEPTIVE SOMATIC SENSORY SIGNALS INTO THE CENTRAL NERVOUS SYSTEM

Either all or almost all sensory information from the somatic segments of the body enters the spinal cord through the dorsal roots of the spinal nerves. However, from this point to the brain the sensory signals are carried through one of two alternate sensory pathways: (1) the *dorsal column–lemniscal system* and (2) the *anterolateral system*. These two systems come together again at the level of the thalamus.

The dorsal column–lemniscal system, as its name implies, carries signals mainly in the *dorsal columns* of the cord and then upward through the brain stem to the thalamus by way of the *medial lemniscus*. On the other hand, signals of the anterolateral system, after originating in the dorsal horns of the spinal gray matter, cross to the opposite side of the cord and ascend through the cord in the anterior and lateral white columns to terminate at all levels of the lower brain stem and also in the thalamus.

The dorsal column–lemniscal system is composed of large, myelinated nerve fibers that transmit signals to the brain at velocities of 30 to 110 meters per second whereas the anterolateral system is composed of much smaller myelinated fibers (averaging 4 microns in diameter) that transmit signals at velocities ranging between 8 and 40 meters per second.

Another difference between the two systems is that the dorsal column–lemniscal system has a very high degree of spatial orientation of the nerve fibers with respect to their origin on the surface of the body, whereas the anterolateral system has a much smaller degree of spatial orientation, with some fibers seeming to have very little orientation at all.

These differences immediately characterize the types of sensory information that can be transmitted by the two systems. First, sensory information that must be transmitted rapidly and with temporal fidelity is transmitted in the dorsal column–lemniscal system, while that which does not need to be transmitted rapidly is transmitted mainly in the anterolateral system. Second, those sensations that detect fine gradations of intensity are transmitted in the dorsal column–lemniscal system, while those that lack the fine gradations are transmitted in the anterolateral system. And, third, sensations that are discretely localized to exact points in the body are transmitted in the dorsal system, while those transmitted in the anterolateral system can be localized much less exactly. On the other hand, the anterolateral system has a special capability that the dorsal system does not have: the ability to transmit a broad spectrum of sensory modalities—pain, warmth, cold, and crude tactile sensations; the dorsal system is limited to mechanoreceptive sensations alone. With this differentiation in mind we can now list the types of sensations transmitted in the two systems.

The Dorsal Column–Lemniscal System

1. Touch sensations requiring a high degree of localization of the stimulus.

2. Touch sensations requiring transmission of fine gradations of intensity.
3. Phasic sensations, such as vibratory sensations.
4. Sensations that signal movement against the skin.
5. Position sensations.
6. Pressure sensations having to do with fine degrees of judgment of pressure intensity.

The Anterolateral System

1. Pain.
2. Thermal sensations, including both warm and cold sensations.
3. Crude touch and pressure sensations capable of only crude localizing ability on the surface of the body and having little capability for intensity discrimination.
4. Tickle and itch sensations.
5. Sexual sensations.

TRANSMISSION IN THE DORSAL COLUMN–LEMNISCAL SYSTEM

ANATOMY OF THE DORSAL COLUMN–LEMNISCAL SYSTEM

On entering the spinal cord from the spinal nerve dorsal roots, the large myelinated fibers from specialized mechanoreceptors first pass medially into the lateral margin of the dorsal white columns. However, almost immediately each of these fibers divides to form a *medial* and a *lateral branch*, as illustrated by the medial dorsal root fiber in Figure 49–2. The medial branch turns upward in the dorsal column and proceeds by way of the dorsal column pathway to the brain

The lateral branch proceeds laterally and anteriorly in the same segment of the cord and divides many times to give off multiple terminals to almost all parts of the intermediate and anterior portions of the cord gray matter. These terminals serve two purposes: (1) Some of them elicit local spinal cord reflexes, which will be

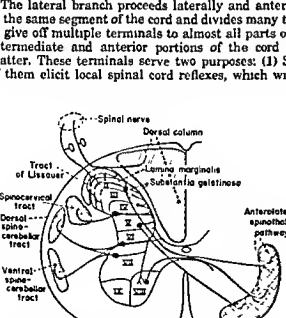


Figure 49–2. Cross-section of the spinal cord showing the anatomical laminae I through IX of the cord gray matter and the ascending sensory tracts in the white columns of the spinal cord.

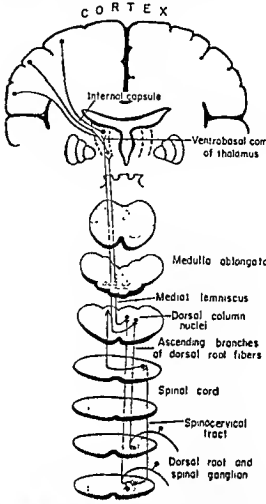


Figure 49–3. The dorsal column and spinocervical pathways for transmitting critical types of tactile signals. (Modified from Ranson and Clark. Anatomy of the Nervous System. Philadelphia, W B Saunders Company, 1959.)

discussed in Chapter 51 (2) Others excite relay neurons that give rise to the spinocerebellar and spinocervical tracts. We will discuss the spinocerebellar tracts in Chapter 53 in relation to the function of the cerebellum. The spinocervical tract is an accessory pathway for transmission of somatosensory information upward in the spinal cord and functions in close association with the dorsal column–lemniscal system, as we shall discuss subsequently.

Anatomy of the Dorsal Column Pathway. Note in Figure 49–3 that the nerve fibers entering the dorsal columns pass up these columns to the medulla, where they synapse in the *dorsal column nuclei* (the *cuneate* and *gracile nuclei*) From here, *second order neurons* decussate immediately to the opposite side and then pass upward to the thalamus through bilateral pathways called the *medial lemnisci*. Each medial lemniscus terminates in a *ventrobasal complex of nuclei* which includes the ventral posterior lateral nucleus of the thalamus and the posterior nuclear group located slightly farther posteriorly. In its pathway through the brain stem, the medial lemniscus is joined by additional fibers from the *main sensory nucleus of the trigeminal nerve* and from the *upper portion of its descending nucleus*; these fibers subserve the same sensory functions for the head that the dorsal column fibers subserve for the body.

From the ventrobasal complex, *third order nerve fibers*

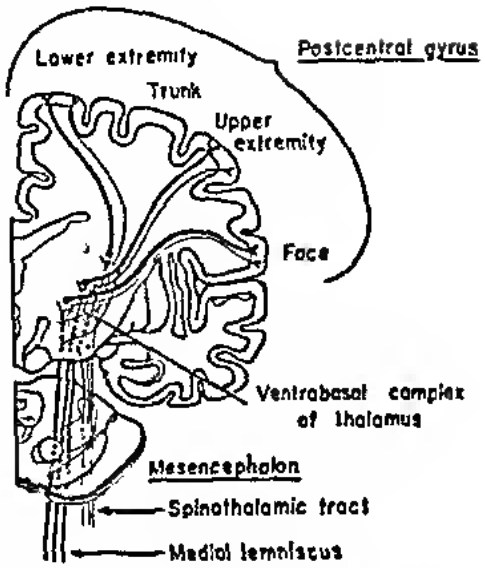


Figure 49–4. Projection of the dorsal column–lemniscal system from the thalamus to the somatic sensory cortex (Modified from Brodal Neurological Anatomy in Relation to Clinical Medicine New York, Oxford University Press, 1969)

project, as shown in Figure 49–4, mainly to the *postcentral gyrus* of the *cerebral cortex*, which is called *somatic sensory area I* In addition, fewer fibers project to the lowermost lateral portion of each parietal lobe, an area called *somatic sensory area II*

Anatomy of the Spino-Cervical Pathways. Until recently it was believed that all the dorsal column–lemniscal system signals were trasmitted only through the dorsal columns However, complete transection of the dorsal column does not cause total loss of transmission in this system Instead, many signals that can partially substitute for the dorsal columns can still be transmitted through the *spinocervical pathway*

The anatomy of the spinocervical pathway is much less well known than that of the dorsal column pathway However, many collateral terminals from the large sensory fibers synapse in lamina IV and adjacent laminae of the cord gray matter and give rise to second order fibers that enter the dorsal portion of the *lateral white column* and ascend in the spinocervical tract to the cervical region of the cord or even to the medulla Here the fibers again synapse either in the dorsal horn of the cord or in medullary nuclei adjacent to or part of the *dorsal column nuclei* Then third order neurons decussate to the opposite side and pass along with the second order neurons of the dorsal column pathway upward to the thalamus through the *medial lemnisci*. Thus, the pathway within the brain parallels that of the dorsal column–lemniscal pathway

Separation of Sensory Modalities Between the Dorsal Column–Lemniscal Pathway and the Spinocervical Pathway

The way in which the nervous system distinguishes among the different modalities of sensation is to transmit the signals of different modalities to separate central nervous system areas. This separation of modality begins in the spinal cord It has already been noted that only mechanoreceptor modalities are transmitted in the dorsal column–lemniscal system whereas all the other somatic modalities are transmitted in the anterolateral system However, even within the dorsal column–lemniscal system there is an additional degree of separation The dorsal column tract transmits mainly signals from rapidly adapting sensory receptors For instance, it is only through the dorsal columns that signals are transmitted from the extremely rapidly adapting pacinian corpuscles. Also, most of the signals from the Meissner's corpuscles and from the hair receptors, both of which are rapidly adapting receptors, are transmitted through this pathway

On the other hand, some of the more slowly adapting signals from the Merkel's discs, from the deep tissue Ruffini end-organs, and from the slowly adapting Ruffini position sense receptors of the joint capsules seem to be transmitted through the spinocervical pathway.

Spatial Orientation of the Nerve Fibers in the Dorsal Column–Lemniscal System

One of the distinguishing features of the dorsal column–lemniscal system is a distinct spatial orientation of nerve fibers from the individual parts of the body that is maintained throughout. For instance, in the dorsal columns, the fibers from the lower parts of the body lie toward the center, while those that enter the spinal cord at progressively higher segmental levels form successive layers laterally

The spatial orientation in the spinocervical pathway is less well known. However, stimulation experiments of single fibers within this pathway have shown that the sensory signals diverge very little, indicating a high degree of spatial orientation in this pathway as well

In the thalamus, the distinct spatial orientation is still maintained, with the tail end of the body represented by the most lateral portions of the ventrobasal complex and the head and face represented in the medial component of the complex However, because of the crossing of the medial lemnisci in the medulla, the left side of the body is represented in the right side of the thalamus, and the right side of the body is represented in the left side of the thalamus.

THE SOMATIC SENSORY CORTEX

The area of the cerebral cortex to which the sensory signals are projected is called the *somatic sensory cortex*. In the human being, this area lies mainly in the anterior portions of the parietal lobes. Two distinct and separate areas are known to receive direct afferent nerve fibers from the relay nuclei of the thalamus; these, called *somatic sensory area I* and *somatic sensory area II*, are illustrated in Figure 49–5. However, somatic sensory area I is so much more important to the sensory functions of the body than is somatic sensory area II that in popular usage, the term somatic sensory cortex is almost always used to mean area I exclusive of area II. Yet, to keep these two areas we will henceforth refer to them sepa-

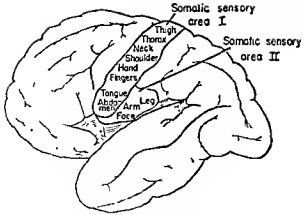


Figure 49–5. The two somesthetic cortical areas, somatic sensory areas I and II.

rately as somatic sensory area I and somatic sensory area II.

Projection of the Body in Somatic Sensory Area I. Somatic sensory area I lies in the postcentral gyrus of the human cerebral cortex (in Brodmann areas 3, 1, and 2 immediately behind the central sulcus—sea map of the cerebral cortex in Figure 54–1 in Chapter 54). A distinct spatial orientation exists in this area for reception of nerve signals from the different areas of the body. Figure 49–6 illustrates a cross-section through the brain at the level of the postcentral gyrus, showing the representations of the different parts of the body in separate regions of somatic sensory area I. Note, however, that each side of the cortex receives sensory information exclusively from the opposite side of the body (with the exception of a small amount of sensory information from the same side of the face).

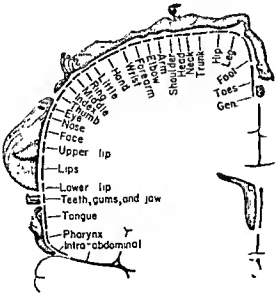


Figure 49–6. Representation of the different areas of the body in the somatic sensory area I of the cortex. (From Penfield and Rasmussen. Cerebral Cortex of Man. A Clinical Study of Localization of Function. New York, Macmillan Company, 1968.)

Some areas of the body are represented by large areas in the somatic cortex—the lips by far the greatest of all, followed by the face and thumb—whereas the entire trunk and lower part of the body are represented by relatively small areas. The sizes of these areas are directly proportional to the number of specialized sensory receptors in each respective peripheral area of the body. For instance, a great number of specialized nerve endings are found in the lips and thumb, while only a few are present in the skin of the trunk.

Note also that the head is represented in the most lateral portion of somatic sensory area I, whereas the lower part of the body is represented in the medial portion of the area.

Somatic Sensory Area II. The second cortical area to which somatic afferent fibers project, somatic sensory area II, is a much smaller area that lies posterior and inferior to the lateral end of somatic sensory area I, as shown in Figure 49–5. The degree of localization of the different parts of the body is very poor in this area compared with somatic sensory area I. The face is represented anteriorly, the arms centrally, and the legs posteriorly.

So little is known about the function of somatic sensory area II that it cannot be discussed intelligently. It is known that signals enter this area from both sides of the body and also from other sensory areas of the brain, such as visual and auditory signals. Also, stimulation of somatic sensory area II in some instances causes complex body movements, for which reason it possibly plays a role in sensory control of motor functions.

The Layers of the Somatic Sensory Cortex and Their Function

The cerebral cortex contains *six* separate layers of neurons, beginning with layer I next to the surface and extending progressively deeper to layer VI. As would be expected, the neurons in each layer perform functions different from those in other layers. Some of these functions are:

1. The incoming sensory signal excites mainly neuronal layer IV first; then the signal spreads toward the surface of the cortex and also toward the deeper layers.

2. Layers I and II receive a diffuse, nonspecific input from the reticular activating system that can facilitate the whole brain at once; this system will be described in Chapter 55. This input perhaps controls the overall level of excitability of the cortex.

3. The neurons in layers II and III send axons to other closely related portions of the cerebral cortex.

4. The neurons in layers V and VI send axons to more distant parts of the nervous system. Those in layer V are generally larger and project to more distant areas. For instance, many of these pass all the way into the brain stem and spinal cord to provide control signals to these areas. From layer VI, especially large numbers of axons extend to the thalamus, providing feedback signals from the cerebral cortex to the thalamus.

Representation of the Different Sensory Modalities in the Somatic Sensory Cortex—The Vertical Columns of Neurons

Functionally, the neurons of the somatic sensory cortex are arranged in vertical columns extending all the way through the six layers of the cortex, each column having a diameter of 0.33 to 1 mm and containing some 100,000 neuronal cell bodies. Each of these columns serves a single specific sensory modality, some responding to stretch receptors around joints, some to stimulation of tactile hairs, others to discrete localized pressure points on the skin, and so forth. Furthermore, the columns for the different modalities are interspersed among each other. At layer IV, where the signals first enter the cord, the columns of neurons function almost entirely separately from each other. However, at other levels of the columns interections occur that allow beginning analysis of the meanings of the sensory signals.

In the most anterior portion of the postcentral gyrus, located deep in the central sulcus in Brodmann area 3a (see Chapter 54 for a map of Brodmann's areas), a disproportionately large share of the vertical columns respond to muscle, tendon, or joint stretch receptors. The signals from these in turn impact mainly on the motor cortex located immediately anterior to the central sulcus and help control muscle function. As one proceeds more posteriorly in somatic sensory cortex I, more and more of the vertical columns respond to the slowly adapting cutaneous receptors, and then still farther posteriorly greater numbers of the columns are sensitive to deep pressure.

In the most posterior portion of somatic sensory area I, about 6 per cent of the vertical columns respond only when a stimulus moves across the skin in a particular direction. Thus, this is a still higher order of interpretation of sensory signals; and the process becomes even more complex still farther posteriorly in the parietal cortex, which is called the *somatic association area* as we shall discuss subsequently.

Functions of Somatic Sensory Area I

The functional capabilities of different areas of the somesthetic cortex have been determined by selective excision of the different portions. Widespread excision of somatic sensory area I causes loss of the following types of sensory judgment:

1. The person is unable to localize discretely the different sensations in the different parts of the body. However, he or she can localize these sensations very crudely, such as to a particular hand, which indicates that the thalamus or parts of the cerebral cortex not normally considered to be concerned with somatic sensations can perform some degree of localization.
2. He is unable to judge critical degrees of pressure against his body.
3. He is unable to judge exactly the weights of objects.
4. He is unable to judge shapes or forms of objects. This is called *astereognosis*.
5. He is unable to judge texture of materials, for this type of judgment depends on highly critical sensations caused by movement of the skin over the surface to be judged.

Note in the list that nothing has been said about loss of pain and temperature sense. However, in the absence of somatic sensory area I, the appreciation of these sensory modalities may be altered either in quality or in intensity. But more important, the pain and temperature sensations that do occur are poorly localized, indicating that both pain and temperature localization probably depend mainly upon simultaneous stimulation of tactile stimuli that use the topographical map of the body in somatic sensory area I to localize the source.

SOMATIC ASSOCIATION AREAS

Brodmann areas 5 and 7 of the cerebral cortex, which are located in the parietal cortex behind somatic sensory area I and above somatic sensory area II, play important roles in deciphering the sensory information that enters the somatic sensory areas. Therefore, these areas are called the *somatic association areas*.

Electrical stimulation in the somatic association area can occasionally cause a person to experience a complex somatic sensation, sometimes even the "feeling" of an object such as a knife or a ball. Therefore, it seems clear that the somatic association area combines information from multiple points in the somatic sensory area to decipher its meaning. This also fits with the anatomical arrangement of the neuronal tracts that enter the somatic association area, for it receives signals from (a) somatic sensory area I, (b) the ventrobasal complex of the thalamus, (c) other areas of the thalamus which themselves receive input from the ventrobasal complex, (d) the visual cortex, and (e) the auditory cortex.

Effect of Removing the Somatic Association Area—Amorphosynthesis. When the somatic association area is removed, the person especially loses the ability to recognize complex objects and complex forms by the process of feeling them. In addition, the person loses most of the sense of form of his or her own body. An especially interesting fact is that loss of the somatic association area on one side of the brain causes the person sometimes to be oblivious of the opposite side of the body—that is, to forget that it is there. Likewise, when feeling objects, the person will tend to feel only one side of the object and to forget that the other

side even exists. This complex sensory deficit ia called *omorphosynthesis.*

CHARACTERISTICS OF TRANSMISSION IN THE DORSAL COLUMN–LEMNISCAL SYSTEM

Faithfulness of Transmission. The most important functional characteristic of the dorsal column–lemniscal system is its *faithfulness* of transmission. Thot is, each time o point in the periphery is stimulated, o signol ordinarily is transmitted all the way to the somatic sensory cortex. Also, if this peripherol stimulus increases in intensity, the intensity of the signal at the cerebral cortex increases opproximately proportionotely. And, finally, when o discrete areo of the body is stimulated, the signol from this area is transmitted to a discrete area of the cerebral cortex. Thus, the dorsal system is adequotely organized for transmission of accurate information from the periphery to the sensorium. Furthermore, the responsiveness of this system can be altered only moderately by stimuli from other areas of the nervous system, and it is not depressed to a significant extent even by moderate degrees of geoeral anesthesia; this is quite different for most portions of the anterolateral sensory system.

Baslo Neuronal Circuit and Dlscharge Pattern in tho Dorsal Column–Lemniscal System. The lower part of Figure 49–7 illustrates the basic organization of the neuronal circuit of the dorsal column pathway, showing that at each synaptic stags divergence occurs. However, the upper part of the figurs shows that a single receptor stimulus on the skin does not cause all the cortical neurons with which that receptor connects to discharge at the same rate. Instead, the cortical neurons that discharge to the greatest extent are those in a central part of the cortical "field" for each respective receptor. Thus, a weak stimulus causes only the ceotralmost neurons to fire. A stronger stimulus causes still more neuroos to fire, but those in the center still discharge at a considerably more rapid rate than do those farther away from the center.

Two-Point Discriminotion. A method frequently used to test tactile capabilities is to determine a person's so-called "two-point discriminatory ability." In this test, two needles are pressed against tbe skio, and the subject determines whether two points of stimulus sre felt or one point. On the tips of tbe fingers a person can distinguish two separate points even when the needles are as close together as 1 to 2 mm. However, on the person's bsck, the needles must usually be as far apart as 30 to 70 mm before one can detect two separate points. The reasoo for this is thst there are many specialized tactile receptors in the tips of the fingers in comparison with a

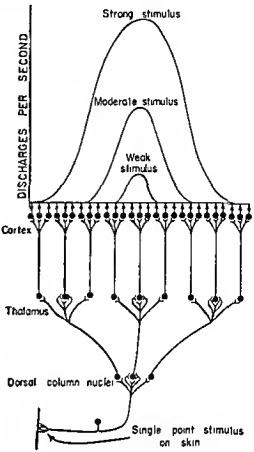


Figure 49–7. Transmission of pinpoint stimulus signal to the cortex.

small number in the skin of the back. Referring back to Figure 49–6, ws cao see also that the portions of the body that have a high degree of two-point discrimination bave a correspondingly large cortical repr ssentation io somatic sensory area I.

Figure 49–8 illustrates the mechanism by which the dorsal column pathway, and the other sensory pathways as well but to a less critical degree, transmit two-point discriminatory informatioo. This figure shows two adjacent points on the skin that are strongly stimulated, and it also shows the small area of the somatic sensory cortex (greatly eolarged) that is excited by signals from the two stimulated points. The solid black curve shows the spatisl pattern of corticsl excitation when both skin points are stimulsted simultaneously. Note that the resultaot zooe of excitation bas two separate peaks. It is these two peaks separated by a valley that allow the sensory cortex to detect the preseoce of two stimulatory points rather than a single point. However, the capability of the seosorium to distinguish between two points of stimulation is strongly influeoced by another mechanism, the mechanism of *lateral inhibition,* as explained io the following section.

Increase in Cootrast in the Perceived

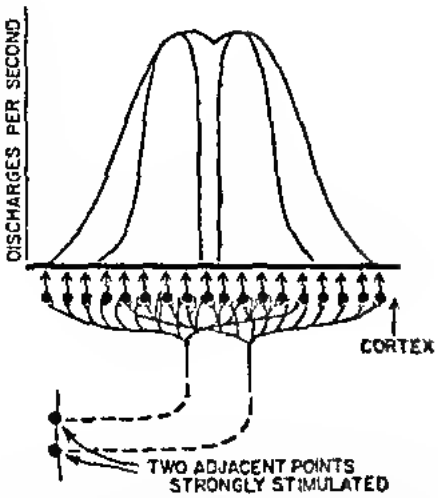


Figure 49–8. Transmission of signals to the cortex from two adjacent pinpoint stimuli. The solid black curve represents the pattern of cortical stimulation without "surround" inhibition, and the two colored curves represent the pattern with "surround" inhibition

Spatial Pattern Caused by Lateral Inhibition. In Chapter 47 it was pointed out that contrast in sensory patterns is increased by inhibitory signals transmitted laterally in the sensory pathway. This effect was illustrated in Figure 47–9 of that chapter

In the case of the dorsal column system, an excited sensory receptor in the skin transmits not only excitatory signals to the somatic sensory cortex but also inhibitory signals laterally to adjacent fiber pathways. These inhibitory signals occur at each synaptic level, for instance in tha dorsal column nuclei and in the ventrobasal complex of the thalamus, and they help block lateral spread of the excitatory signal, a process called *lateral inhibition* or *surround inhibition*. As a result, the peak of excitation stands out, and much of the surrounding diffuse stimulation is blocked. This effect is illustrated by the two colored curves in Figure 49–8, showing complete separation of the peaks when the intensity of the surround inhibition is very great. Obviously, this mechanism accentuates the contrast between the areas of peak stimulation and the surrounding areas, thus greatly increasing the contrast or sharpness of the perceived spatial pattern

Transmission of Rapidly Chonging and Repetitive Sensations. The dorsal column system is of particular value for apprising the sensorium of rapidly changing peripheral conditions. This system can "follow" changing stimuli up to at least 400 cycles per second and can "detect" changes as high as 700 cycles. Obviously, very rapid response to a sudden stimulus on the skin allows one to direct attention immediately to any point of contact, which, in turn, allows one to make necessary corrections before damage can be done.

Vibratory Sensation. Vibratory signals are rapidly repetitive and can be detected as vibration up to 700 cycles per second. These signals are transmitted only in the dorsal column pathway and not in the more slowly conducting pathways For this reason, application of vibration with a tuning fork to different peripheral parts of the body is an important tool used by the neurologist for testing functional integrity of the dorsal columns.

THE POSITION SENSE

The term position sense can be divided into two subtypes: (1) *static position*, which means conscious recognition of the orientation of the different parts of the body with respect to each other, and (2) *kinesthesia*, which means conscious recognition of rates of movement of the different parts of the body. The position sensations are transmitted to the sensorium almost entirely through the dorsal column–lemniscal system. However, special features of the position sense require further explanation.

The Position Sense Receptors. Sensory information from many different types of receptors is used to determine both static position and kinesthesia These include the extensive sensory endings in the joint capsules and ligaments, the receptors in the skin and deep tissues near the joints, and probably also the muscle spindles that detect muscle stretch, as we shall discuss in Chapter 51.

Three major types of nerve endings have been described in the joint capsules and ligaments about the joints. (1) By far the most abundant of these are spray-type *Ruffini endings*, one of which was illustrated in Figure 48–1 of Chapter 48. These endings are stimulated strongly when the joint is suddenly moved; they adapt slightly at first but then transmit a steady signal thereafter. (2) A second type of ending resembling the stretch receptors found in muscle tendons (called *Golgi tendon receptors*) is found particularly in the ligaments about the joints. Though far less numerous than the Ruffini endings, they have essentially the same response properties. (3) A few *pacinian corpuscles* are also found in the tissues around the joints. These adapt extremely rapidly and presumably help to detect *rate of rotation* at the joint.

Detection of Static Position by the Joint Receptors. Figure 49–9 illustrates the excitation of seven different nerve fibers leading from separate joint receptors in the capsule of a cat's knee joint. Note that at 180 degrees of joint rotation one of the receptors is stimulated; then at 150 degrees still another is stimulated; at 140 degrees two are stimulated, and so forth. The information from these joint receptors continually apprises the central nervous system of the momentary rotation of the joint. That is, the rotation determines *which* receptor is stimulated and how much it is stimulated, and from this the brain knows how far the joint is bent.

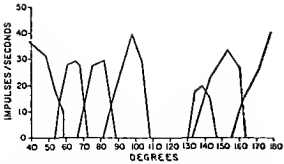


Figure 49–9. Responses of seven different nerve fibers from knee joint receptors in a cat at different degrees of rotation (Modified from Skoglund *Acta Physiol Scand.* Suppl 124, 361, 1956)

Detection of Rate of Movement (Kinesthesia) at the Joint. Because pacinian corpuscles are especially adapted for detecting movement of tissues, it is tempting to suggest that rate of movement at the joints is detected by the pacinian corpuscles. However, the number of pacinian corpuscles in the joint tissues is small, for which reason rate of movement at the joint is probably detected mainly in the following way: The Ruffini and Golgi endings in the joint tissues are stimulated very strongly at first by the joint movement, but within a fraction of a second this strong level of stimulation fades to a lower, steady state rate of firing. Nevertheless, this early overshoot in receptor stimulation is directly proportional to the rate of joint movement and is believed to be the signal used by the brain to discern the rate of movement.

Transmission of Position Sense Signals in the Dorsal Column–Lemniscal Pathway. In the past it had been believed that essentially all the position sense signals were transmitted in the dorsal columns. However, on the basis of studies in animals, there is now much reason to believe that some of these signals are transmitted in the spinocervical pathway as well. This is probably especially true of the static position signals, that is, those signals that are slowly adapting and that transmit information about the relative positions of the different parts of the body when the person is not moving.

Processing of Position Sense Information in the Dorsal Column–Lemniscal Pathway. Despite the faithfulness of transmission of signals from the periphery to the sensory cortex in the dorsal column–lemniscal system, there nevertheless seems to be some processing of position sense signals before they reach the cerebral cortex. For instance, Figure 49–9 showed that individual joint receptors are stimulated maximally at specific degrees of rotation of the joint, with the intensity of stimulation decreasing on either side of the maximal point for each receptor. However, the static position signal for joint rotation is quite different at the level of the ventrobasal complex of the thalamus, as can be seen by referring to Figure

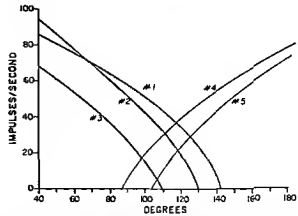


Figure 49–10. Typical responses of five different neurons in the knee joint receptor field of the ventrobasal complex when the knee joint is moved through its range of motion (The curves were constructed from data in Mountcastle et al. *J Neurophysiol*, 26 807, 1963)

49–10. This figure shows that the ventrobasal neurons that respond to the joint rotation signal are of two types: (1) those that are maximally stimulated when the joint is at full rotation and (2) those that are maximally stimulated when the joint is at minimal rotation. In each case, as the degree of rotation changes, the rate of stimulation of the neuron either decreases or increases, depending on the direction in which the joint is being rotated. Furthermore, the intensity of neuronal excitation changes over angles of 40 to 60 degrees of angulation in contrast to 20 to 30 degrees for the individual receptors, as was illustrated in Figure 49–9. Thus, the signals from the individual joint receptors have been integrated in the space domain by the time they reach the thalamic neurons, giving a progressively stronger signal as the joint moves in only one direction rather than giving a peaked signal as occurs in stimulation of individual receptors

TRANSMISSION IN THE ANTEROLATERAL SYSTEM

It was pointed out earlier in the chapter that the anterolateral system transmits sensory signals that do not require highly discrete localization of the signal source and also do not require discrimination of fine gradations of intensity. These include pain, heat, cold, crude tactile, tickle and itch, and sexual sensations. In the following chapter pain and temperature sensations will be discussed; the present chapter is concerned principally with transmission of the tactile sensations.

ANATOMY OF THE ANTEROLATERAL PATHWAY

The anterolateral fibers originate mainly in laminae I, IV, V, and VI in the dorsal horns where many of the dorsal root sensory nerve fibers terminate after entering

the cord (see Figure 49–2). Then, as illustrated in Figure 49–11, the fibers cross in the anterior commissure of the cord to the opposite anterolateral white column where they turn upward toward the brain. These fibers ascend rather diffusely throughout the anterolateral columns. However, anatomical studies have shown that there is at least a partial differentiation of this pathway into an anterior division, called the *anterior spinothalamic tract*, and a lateral division, called the *lateral spinothalamic tract*. Also encompassed in the anterolateral pathway are a *spinoreticular pathway* (to the reticular substance of the brain stem) and a *spinotectal tract* (to the tectum of the mesencephalon). However, it has been difficult to make these differentiations using electrical recording techniques.

The upper terminus of the anterolateral pathway is mainly twofold: (1) throughout the *reticular nuclei of the brain stem*, and (2) in two different nuclear complexes of the thalamus, the *ventrobasal complex* and the *intralaminar nuclei*. In general, the tactile signals are transmitted mainly into the ventrobasal complex, terminating in the same thalamic nuclei as the dorsal column–lemniscal system, and this is probably also true for the temperature signals. From here, the tactile signals are transmitted to the somatosensory cortex along with the signals from the dorsal columns. On the other hand, only part of the pain signals project to this complex. Instead, most of these enter the reticular nuclei of the brain stem and intralaminar nuclei of the thalamus, as

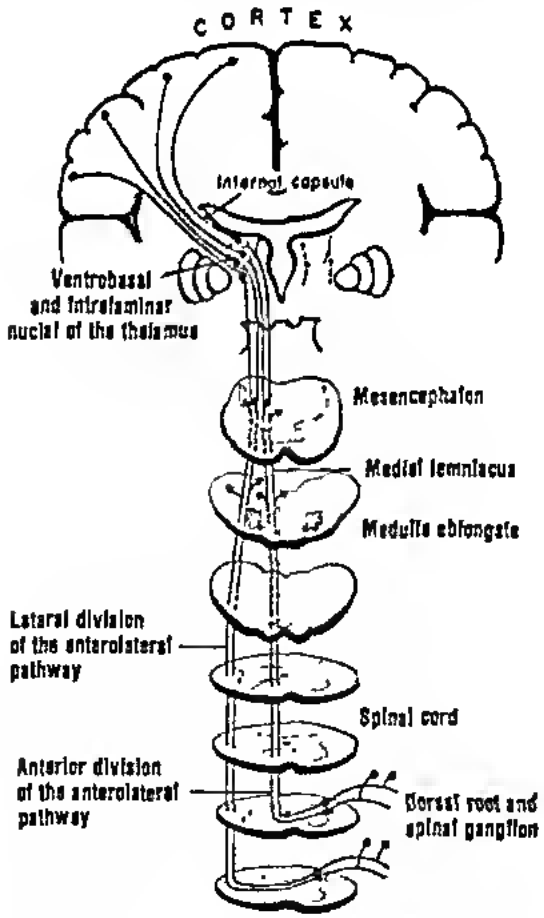


Figure 49–11. The anterior and lateral divisions of the anterolateral pathway.

will be discussed in greater detail in the following chapter.

Characteristics of Transmission in the Anterolateral Pathway. In general, the same principles apply to transmission in the anterolateral pathway as in the dorsal column–lemniscal system except for the following differences: (a) the velocities of transmission are only one half to one third those in the dorsal column–lemniscal system, ranging between 8 and 40 meters per second; (b) the degree of spatial localization of signals is poor, especially in the pain pathways; (c) the gradations of intensities are also far less acute, most of the sensations being recognized in 10 to 20 gradations of strength rather than as many as 100 gradations for the dorsal column system; and (d) the ability to transmit rapidly repetitive sensations is poor.

Thus, it is evident that the anterolateral system is a cruder type of transmission system than the dorsal column–lemniscal system. Even so, certain modalities of sensation are transmitted in this system only and not at all in the dorsal column–lemniscal system. These are pain, thermal, tickle and itch, and sexual sensations in addition to crude touch and pressure.

SOME SPECIAL ASPECTS OF SENSORY FUNCTION

Function of the Thalamus in Somatic Sensation. When the somatosensory cortex of a human being is destroyed, that person loses most critical tactile sensibilities, but a slight degree of crude tactile sensibility does return. Therefore, it must be assumed that the thalamus (as well as other lower centers) has a slight ability to discriminate tactile sensation even though the thalamus normally functions mainly to relay this type of information to the cortex.

On the other hand, loss of the somatosensory cortex has little effect on one's perception of pain sensation and only a moderate effect on the perception of temperature. Therefore, there is much reason to believe that the thalamus and other associated basal regions of the brain play perhaps the dominant role in discrimination of these sensibilities; it is interesting that these sensibilities appeared very early in the phylogenetic development of animalhood while the critical tactile sensibilities were a late development.

Cortical Control of Sensory Sensitivity. Almost all sensory information that enters the cerebrum, except that from the olfactory system, but including sensory information from the eyes, the ears, the taste receptors, and all the somatic receptors, is relayed through one or another of the thalamic nuclei. Furthermore, the conscious brain is capable of directing its attention to different segments of the sensory system. This function is believed to be mainly achieved through facilitation or inhibition of the cortical receptive areas, perhaps as a result of signals from the thalamus.

But, in addition, "corticofugal" signals are transmitted from the cortex to the lower relay stations in the sensory pathways to *inhibit* transmission. For instance, corticofugal pathways control the sensitivity of the synapses

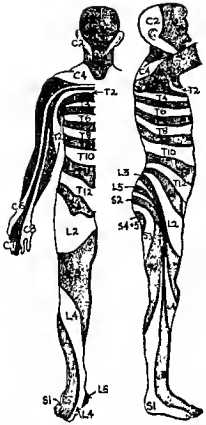


Figure 49–12. The dermatomes. (Modified from Grinker and Sahs. Neurology Charles C Thomas, 1966)

at all levels of the sensory pathways—in the thalamus, in the brain stem reticular nuclei, in the dorsal column nuclei, and especially in the dorsal horn relay station of the anterolateral system. Also, similar inhibitory mechanisms are known for the visual, auditory, and olfactory systems, which are discussed in later chapters Each corticofugal pathway begins in the cortex where the sensory pathway that it controls terminates. Thus, a feedback control loop exists for each sensory pathway.

Obviously, corticofugal control of sensory input could allow the cerebral cortex to alter the threshold for different sensory signals. Also, it might help the brain focus its attention on specific types of information, which is an important and necessary characteristic of nervous system function.

Automatic Gain Control in the Sensory Pathways, and Enhancement of Contrast. One of the byproducts of corticofugal inhibition is believed to be automatic gain control in many if not most of the sensory pathways. That is, when excess sensory signals pass to the brain, recurrent inhibition back to the cord decreases the sensitivity of the sensory pathway. This obviously prevents overloading of the pathway with signals. Also, it helps to preserve contrast in the perception of sensory signals because an excess of signals in the pathway would deluge both the relay nuclei and the cerebral cortex with such a high degree of excitation that it would be impossible to distinguish the highlights from the chaff

Segmental Fields of Sensation—The Dermatomes. Each spinal nerve innervates a "segmental field" of the skin called a *dermatome*. The different dermatomes are illustrated in Figure 49–12. However, these are shown as if there were distinct borders between the adjacent dermatomes, which is far from true because much overlap exists from segment to segment. Indeed, because of the great overlap, the posterior roots from an entire segment of the spinal cord can be destroyed without causing significant loss of sensation in the skin.

Figure 49–12 shows that the anal region of the body lies in the dermatome of the most distal cord segment. In the embryo, this is the tail region and is the most distal portion of the body. The legs develop from the lumbar and upper sacral segments rather than from the distal sacral segments, which is evident from the dermatomal map. Obviously, one can use a dermatomal map such as that illustrated in Figure 49–12 to determine the level in the spinal cord at which various cord injuries may have occurred when the peripheral sensations are disturbed.

REFERENCES

See bibliographies of Chapters 48 and 50

50

Somatic Sensations: II. Pain, Visceral Pain, Headache, and Thermal Sensations

Many, if not most, ailments of the body cause pain. Furthermore, the ability to diagnose different diseases depends to a great extent on a doctor's knowledge of the different qualities of pain. For these reasons, the present chapter is devoted mainly to pain and to the physiologic basis of some of the associated clinical phenomena

The Purpose of Pain. Pain is a protective mechanism for the body; it occurs whenever any tissues are being damaged, and it causes the individual to react to remove the pain stimulus. Even such simple activities as sitting for a long time on the ischia can cause tissue destruction because of lack of blood flow to the skin where the skin is compressed by the weight of the body. When the skin becomes painful as a result of the ischemia, the person shifts weight unconsciously A person who has lost the pain sense, such as after spinal cord injury, fails to feel the pain and therefore fails to shift weight This eventually results in ulceration at the areas of pressure unless special measures are taken to move the person from time to time.

THE TWO TYPES OF PAIN AND THEIR QUALITIES—ACUTE PAIN AND SLOW PAIN

Pain has been classified into two different major types: *acute pain* and *slow pain*. Acute pain occurs within about 0.1 second when a pain stimulus is applied, whereas slow pain begins only after a second or more and then increases slowly over a period of many seconds and sometimes even minutes During the course of this chapter we shall see that the conduction pathways for these two types of pain are different and that each of them has specific qualities.

Acute pain is also described by many alternate names, such as *sharp pain*, *pricking pain*, *fast pain*, *electric pain*, and others. This type of pain is felt when a needle is stuck into the skin or when the skin is cut with a knife, and this pain is also felt when the skin is subjected to electric shock Acute, sharp pain is not felt in most of the deeper tissues of the body

Slow pain goes by multiple additional names such as *burning pain*, *aching pain*, *throbbing pain*, *nauseous pain*, and *chronic pain* This type of pain is usually associated with *tissue destruction*. It can become excruciating and can lead to prolonged, unbearable suffering. It can occur both in the skin and in almost any internal tissue or organ.

We will learn later that the acute type of pain is transmitted through type Aδ pain fibers, whereas the slow type of pain results from stimulation of the more primitive type C fibers.

METHODS FOR MEASURING THE PERCEPTION OF PAIN

The intensity of a stimulus necessary to cause pain can be measured in many different ways, but the most used methods have been pricking the skin with a pin at measured pressures, pressing a solid object against a protruding bone with measured force, pinching the skin, or heating the skin with measured amounts of heat. The last method has proved to be especially accurate from a quantitative point of view.

Figure 50–1 illustrates the basic principles of a heat apparatus used for measuring pain threshold An intense light is focused by a large condenser lens onto a black spot painted on the forehead of the subject, and the heat intensity delivered by the light is controlled by a rheostat. In determining the subject's threshold for pain, the intensity of the heat is increased in progressive steps, and the length of time required for the forehead to heat sufficiently to elicit pain is recorded for each heat intensity. These data are then plotted in the form of a "strength-duration curve" to express pain threshold, as shown in Figure 50–2. The lowest intensity of stim-

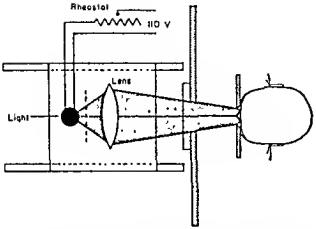


Figure 50–1. Heat apparatus for measuring pain threshold (from Hardy, Wolff, and Goodell. *J. Clin. Invest.* 19 649, 1940)

ulus, illustrated by the dashed line in the figure, that will excite the sensation of pain when the stimulus is applied for a prolonged period of time is called the *pain threshold.*

The Pain Threshold. Figure 50–3 shows graphically the skin temperature at which pain is first perceived by different persons. By far the greatest number perceive pain when the skin temperature reaches almost exactly 45° C, and almost everyone perceives pain before the temperature reaches 47° C. Indeed, measurements in people as widely different as Eskimos, Indians, and whites have shown little differences in their *thresholds for pain.* However, different people do *react* very differently to pain, as is discussed later.

THE PAIN RECEPTORS AND THEIR STIMULATION

Free Nerve Endings as Pain Receptors. The pain receptors in the skin and other tissues are all free nerve endings. They are widespread in the superficial layers of the *skin* and also in certain internal tissues, such as the *periosteum,* the *arterial walls,* the *joint surfaces,* and the *falx* and *tentorium* of the cranial vault. Most of the other deep tissues are not extensively supplied with pain endings but are weakly supplied; nevertheless, any widespread tissue damage can still summate to cause the slow-chronic-aching type of pain in these areas.

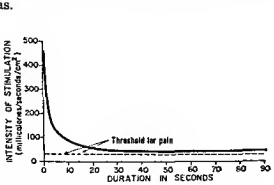


Figure 50–2. Strength-duration curve for depicting pain threshold (from Hardy. *J. Chronic Dis.,* 4 20, 1956)

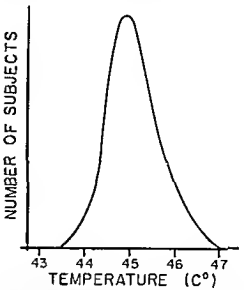


Figure 50–3. Distribution curve obtained from a large number of subjects of the minimal skin temperature that causes pain. (Modified from Hardy. *J. Chronic Dis.,* 4 22, 1956)

Types of Stimuli That Excite Pain Receptors—Mechanical, Thermal, and Chemical. Some pain fibers are excited almost entirely by excessive mechanical stress or mechanical damage to the tissues; these are called *mechanosensitive pain receptors.* Others are sensitive to extremes of heat or cold and therefore are called *thermosensitive pain receptors.* And still others are sensitive to various chemical substances and are called *chemosensitive pain receptors.* Some of the different chemicals that excite the chemosensitive receptors include *bradykinin, serotonin, histamine, potassium ions, acids, prostaglandins, acetylcholine,* and *proteolytic enzymes.*

Though some pain receptors are mainly sensitive to only one of the above types of stimuli, most are sensitive to more than one of the types.

Nonadapting Nature of Pain Receptors. In contrast to most other sensory receptors of the body the pain receptors adapt either not at all or almost not at all. In fact, under some conditions, the threshold for excitation of the pain fibers becomes progressively lower and lower as the pain stimulus continues, thus allowing these receptors to become progressively more activated with time. This increase in sensitivity of the pain receptors is called *hyperalgesia.*

One can readily understand the importance of this failure of pain receptors to adapt, for it allows them to keep the person apprised of a damaging stimulus that causes the pain as long as it persists.

RATE OF TISSUE DAMAGE AS A CAUSE OF PAIN

The average critical temperature of 45° C at which a person first begins to perceive pain is also the temperature at which the tissues begin to be damaged by heat; indeed, the tissues are eventu-

ally completely destroyed if the temperature remains at this level indefinitely. Therefore, it is immediately apparent that pain resulting from heat is closely correlated with the ability of heat to damage the tissues.

Furthermore, the intensity of pain has also been closely correlated with the rate of tissue damage caused by other effects besides heat, whether this be damage caused by bacterial infection, tissue ischemia, tissue contusion, or otherwise.

Special Importance of Chemical Pain Stimuli During Tissue Damage. Extracts from damaged tissues cause intense pain when injected beneath the normal skin. Among the substances in such extracts that are especially painful are *bradykinin, histamine, prostaglandins, acids, excesses of potassium ions, serotonin,* and *proteolytic enzymes,* which are the same substances that are known from electrophysiological data to excite the pain nerve endings. Obviously, many of these substances could cause direct damage to the pain nerve endings, especially the proteolytic enzymes. But some of the other substances, such as bradykinin and some of the prostaglandins, can cause direct extreme stimulation of pain nerve fibers without necessarily damaging them.

Release of the various substances listed above not only stimulates the chemosensitive pain endings but also greatly decreases the threshold for stimulation of the mechanosensitive and thermosensitive pain receptors as well. A widely known example of this is the extreme pain caused by slight mechanical or heat stimuli following tissue damage by sunburn.

Tissue Ischemia as a Cause of Pain. When blood flow to a tissue is blocked, the tissue becomes very painful within a few minutes. And the greater the rate of metabolism of the tissue, the more rapidly the pain appears. For instance, if a blood pressure cuff is placed around the upper arm and inflated until the arterial blood flow ceases, exercise of the forearm muscles can cause severe muscle pain within 15 to 20 seconds. In the absence of muscle exercise, the pain will not appear for three to four minutes. Cessation of blood flow to the skin, in which the metabolic rate is very low, usually does not cause pain for about 20 to 30 minutes.

One of the suggested causes of pain in ischemia is accumulation of large amounts of lactic acid in the tissues, formed as a consequence of the anaerobic metabolism (metabolism without oxygen) that occurs during ischemia. However, it is also possible that other chemical agents, such as bradykinin, proteolytic enzymes, and so forth, are formed in the tissues because of cell damage and that these, rather than lactic acid, stimulate the pain nerve endings.

Muscle Spasm as a Cause of Pain. Muscle spasm is also a very common cause of pain, and it is the basis of many clinical pain syndromes. This pain probably results partially from the direct effect of muscle spasm in stimulating mechanosensitive pain receptors. However, it possibly results also from the indirect effect of muscle spasm in causing ischemia and thereby stimulating chemosensitive pain receptors. The muscle spasm not only compresses the blood vessels and diminishes blood flow but also increases the rate of metabolism in the muscle tissue at the same time, thus making the relative ischemia even greater, creating ideal conditions for release of chemical pain-inducing substances.

THE DUAL PATHWAY FOR TRANSMISSION OF PAIN SIGNALS INTO THE CENTRAL NERVOUS SYSTEM

Even though all pain endings are free nerve endings, these endings utilize two separate pathways for transmitting pain signals into the central nervous system. The two pathways correspond to the two different types of pain, an *acute-sharp pain pathway,* and a *slow-chronic pain pathway.*

The Peripheral Pain Fibers—"Fast" and "Slow" Fibers. The acute sharp pain signals are transmitted in the peripheral nerves to the spinal cord by small type Aδ fibers at velocities of between 6 and 30 meters per second. On the other hand, the slow-chronic type of pain is transmitted by type C fibers at velocities of between 0.5 and 2 meters per second. When the type Aδ fibers are blocked without blocking the C fibers by moderate compression of the nerve trunk, the acute sharp pain disappears. On the other hand, when the type C fibers are blocked without blocking the delta fibers by low concentrations of local anesthetic, the slow-chronic-aching type of pain disappears.

Because of this double system of pain innervation, a sudden onset of painful stimulus gives a "double" pain sensation: a fast, sharp pain followed a second or so later by a slow, burning pain. The sharp pain apprises the person very rapidly of a damaging influence and, therefore, plays an important role in making the person react immediately to remove himself or herself from the stimulus. On the other hand, the slow, burning sensation tends to become more and more painful over a period of time. This sensation gives one the intolerable suffering of long-continued pain.

Dual Processing of the Pain Signals in the Spinal Cord. On entering the spinal cord from the dorsal spinal roots, the pain fibers ascend or descend one to three segments in the *tract of Lissauer* that lies immediately posterior to the dorsal horn of the cord gray matter. Then they terminate on neurons in the dorsal horns. However, here again, there are two systems for processing the pain signals prior to relaying the signals to the brain, as illustrated in Figure 50–4. These are:

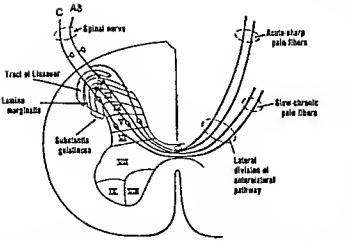


Figure 50–4. Transmission of both "acute-sharp" and "slow-chronic" pain signals into and through the spinal cord on the way to the brain stem.

Processing of the "fast" pain signals: The "fast" type Aδ pain fibers terminate at two points in the dorsal horns, in lamina I (lamina marginalis) and lamina V, as shown by the medialmost dorsal root fiber in the figure. In both of these laminaa the incoming pain fibers excite second order neurons that send long fibers immediately to the opposite side of the cord in the anterior commissure and then upward to tha brain in the lateral division of the anterolateral sensory pathway, as described in tha previous chapter.

Processing of the "slow" signals: The type C fibers that transmit the "slow" pain signals terminate almost entirely in laminae II and III of the dorsal horns, an area called the *substantia gelatinosa* as shown by the lateralmost dorsal root fiber in the figure. Most of the signals then pass through one or more additional short-fiber neurons, eventually terminating mainly in lamina V. Here the last neuron in the series gives rise to long axons, most of which join the fibers from the fast pathway and pass through the anterior commissure to the opposite side of the cord, then upward to the brain via tbe lateral division of the anterolateral sensory pathway. However, a few of these fibers do not cross but instead pass ipsilaterally to the brain.

In the lateral division of the anterolateral pathway, the fast pain fibers and the slow fibers remain differentiated from each other, the fast fibers having diameters of 3 to 5 microns and the slow fibers 1 to 3 microns, and with correspondingly faster transmission in the fast fibers.

Termination of the Fast-Acute Pain Pathway in the Brain Stem and Thalamus. Figure 50–5 illustrates the pain pathways entering the brain stem from the spinal cord. About three quarters to nine tenths of all pain fibers terminate in the reticular formation of the medulla, pons, and mesencephalon. From these areas, higher order neurons are transmitted to the thalamus, hypothalamus, and other areas of the diencephalon and cerebrum.

However, a small proportion of the pain fibers, especially those that transmit the fast-acute type of pain, pass directly to the thalamus and terminate in the ventrobasal complex and posterior nuclear group of the thalamus, along with the sensory fibers from the dorsal column–lemniscal pathway. From here signals are transmitted into other areas of the thalamus and also to the somatic sensory cortex. These signals to the cortex are probably important mainly for localizing the pain, not for interpreting it.

Termination of the Slow-Chronic Pain Fibers in the Brain Stem and Thalamus. In contrast to the fibers of the acute-fast pain pathway, those of the slow-chronic pathway terminate almost entirely in the reticular formation of the hrain stem. However, great numbers of signals are relayed upward through this formation and finally into the intralaminar nuclei of tha thalamus. Both the reticular area of the brain stem and the intralaminar nuclei are parts of the reticular activating system. In Chapter 55 we will discuss in detail the functions of tbis system; briefly, it transmits activating signals into essentially all parts of tha brain, especially upward through the thalamus to all areas of the cerebral cortex and also laterally into tha basal regions of the brain around tha thalamus, including very importantly tbe hypothalamus.

Thus, the slow-burning-aching pain fibers, becausa they do excite the reticular activating system, have a very potent effect in activating essen-

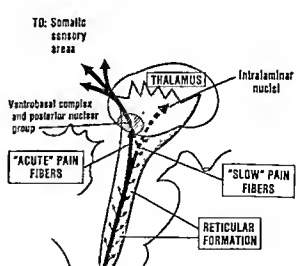


Figure 50–5. Transmission of pain signals into the hindbrain, thalamus, and cortex via the "pricking pain" pathway and the "burning pain" pathway.

tially the entire nervous system—that is, to arouse one from sleep, to create a state of excitement, to create a sense of urgency, and to promote defense and aversion reactions designed to rid the person or animal of the painful stimulus.

The signals that are transmitted through the slow-burning-aching pain pathway can be localized only to very gross areas of the body. Therefore, these signals are designed almost entirely for the single purpose of calling one's attention to injurious processes in the body. They create suffering that is sometimes intolerable. Their gradation of intensity is poor; instead, even weak pain signals can summate over a period of time by a process of temporal summation to create an unbearable feeling even though the same pain for short periods of time may be relatively mild

Function of the Reticular Formation, Thalamus, and Cerebral Cortex in the Appreciation of Pain. Complete removal of the somatic sensory areas of the cerebral cortex does not destroy one's ability to perceive pain. Therefore, it is believed that pain impulses entering only the reticular formation, thalamus, and other lower centers can cause conscious perception of pain. However, this does not mean that the cerebral cortex has nothing to do with normal pain appreciation; indeed, electrical stimulation of the cortical somatic sensory areas causes a person to perceive mild pain in approximately 3 per cent of the stimulations It is believed that the cortex plays an important role in interpreting the quality of pain even though pain perception might be a function of lower centers

Localization of Pain in the Body. Most localization of pain probably results from simultaneous stimulation of tactile receptors along with the pain stimulation However, the acute type of pain, transmitted through type Aδ fibers, can be localized perhaps within 10 to 20 cm of the stimulated area. On the other hand, the slow-burning-aching type of pain, transmitted through C fibers, is localizable only very grossly, perhaps to a major part of the body such as a limb but certainly not to small areas. This is in keeping with the fact that these fibers terminate extremely diffusely in the brain stem and thalamus

Surgical Interruption of Pain Pathways. Often a person has such severe and intractable pain (this often results from rapidly spreading cancer) that it is necessary to relieve the pain To do this the pain pathway can be destroyed at any one of several different points If the pain is in the lower part of the body, a *cordotomy* in the upper thoracic region often relieves the pain for a few weeks to a few months To do this, the spinal cord on the side opposite the pain is sectioned almost entirely through its anterolateral quadrant, which interrupts the anterolateral sensory pathway

Unfortunately, though, the cordotomy is not always successful in relieving the pain for two reasons First, many of the pain fibers from the upper part of the body do not cross to the opposite side of the spinal cord until they have reached the brain so that the cordotomy does not transect these fibers. Second, pain frequently return several months later; indeed, the new pain is often eve more objectionable than the original pain This pai presumably is transmitted in the dorsal half of the cor but the locus and destination of the fiber pathways ar unknown.

An experimental operative procedure to relieve pai is to place lesions in or around the pain-receptive area of the thalamus to block the pain pathway at this poin It has been claimed that destruction of portions of th intralaminar nuclei in the thalamus can relieve th suffering elicited by pain while still leaving intact one' appreciation of "acute" pain, which remains an impo tant protective mechanism.

A PAIN CONTROL ("ANALGESIA") SYSTEM IN THE BRAIN AND SPINAL CORD

The degree to which each person reacts to pain varies tremendously This results partly from the capability of the brain itself to control the degree of input of pain signals to the nervous system by activation of a pain control system, called an *analgesia system*

The enalgesia system is illustrated in Figure 50–6 It consists of three major components (plus other accessory components):

(1) The *periaqueductal gray area* of the mesencephalon and upper pons surrounding the equeduct

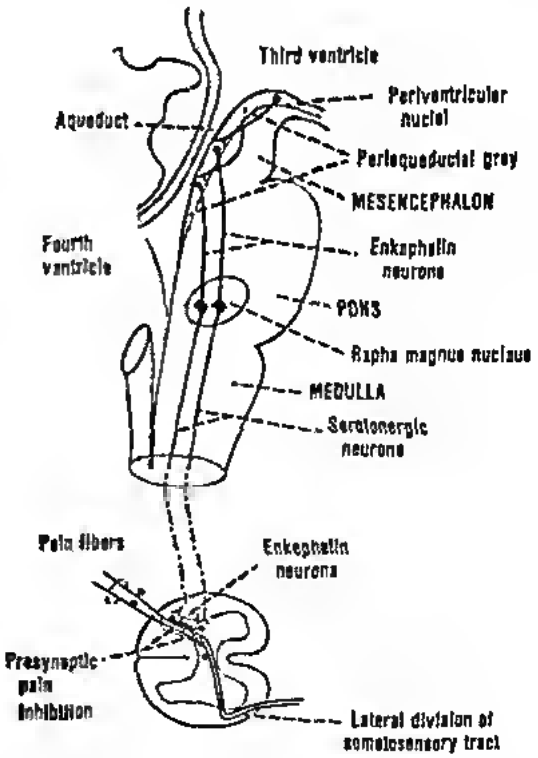


Figure 50–6. The analgesia system of the brain stem and spinal cord, showing inhibition of incoming pain signals at the cord level.

of Sylvius. Neurons from this area send their signals to

(2) the *raphe magnus nucleus*, a thin midline nucleus located in the lower pons and upper medulla. From here the signals are transmitted down the spinal cord to

(3) a *pain inhibitory complex located in the dorsal horns of the spinal cord*. At this point the pain signals can be blocked before they are relayed on to the brain.

Electrical stimulation in either the periaqueductal gray area or in the raphe magnus nucleus can almost completely suppress many very strong pain signals entering by way of the dorsal spinal roots. Also, stimulation of areas at still higher levels of the brain that in turn excite the periaqueductal gray, especially the *periventricular nuclei in the hypothalamus* lying adjacent to the third ventricle and to a lesser extent the *medial forebrain bundle* also in the hypothalamus, can suppress pain though perhaps not quite so much so

Several different transmitter substances are involved in the analgesia system; especially involved are *enkephalin* and *serotonin*. Many of the nerve fibers derived from both periventricular nuclei and the periaqueductal gray area secrete enkephalin at their endings. Thus, as shown in Figure 50–6, the endings of many of the fibers in the raphe magnus nucleus release enkephalin The fibers originating in this nucleus but terminating in the dorsal horns of the spinal cord secrete serotonin at their endings. Tha serotinin in turn acts on still another set of local cord neurons that are believed to secrete enkephalin. The enkephalin, in some way not presently understood, is believed to cause presynaptic inhibition of the incoming pain fibers to laminae I through V in the dorsal horns

Thus, the analgesia system can block pain signals at the initial entry point to the spinal cord. In fact, it can also block many of the local cord reflexes that result from pain signals, especially the withdrawal reflexes that will be described in the following chapter.

It is probable that this analgesia system can also inhibit pain transmission at other points in the pain pathway, especially in the reticular nuclei in the brain stem and in the intralaminar nuclei of the thalamus.

Most important, it must be noted that this pain analgesia system is capable of blocking both the fast-sharp type of pain and also the slow-burning-aching type of pain.

The Brain's Opiate System—The Endorphins and Enkephalins

More than 20 years ago it was discovered that injection of extremely minute quantities of morphine into either the periventricular nucleus around the third ventricle of the diencephalon or into the periaqueductal gray area of the brain stem will cause an extreme degree of analgesia. In subsequent studies, it has now been found that morphine acts at still many other points in the analgesia system, including both in the raphe magnus nucleus and in the dorsal horns of the spinal cord. Because most drugs that alter the excitability of neurons do so by acting on synaptic receptors, it was assumed that the "morphine-receptors" of the analgesia system must in fact be receptors for some morphine-like neurotransmitter that is naturally secreted in the brain. Therefore, 20 years ago an extensive search was set into motion for a natural opiate of the brain. At least nine such opiate-like substances have now been found in different points of the nervous system. Furthermore, multiple areas of the brain have been shown to have opiate receptors, especially the areas in the analgesia system. Among the more important of the opiate substances are:

(a) β-*endorphin*,
(b) *met-enkephalin*,
(c) *leu-enkephalin*, and
(d) *dynorphin*

The two enkephalins are found in the portions of the analgesia system described earlier, and β-endorphin is present both in the hypothalamus and in the pituitary gland. In addition, the pituitary gland contains a large peptide, β-*lipotropin*, that includes within its molecule β-endorphin; and β-endorphin includes within its molecula met-enkephalin and presumably can release this enkephalin slowly over a prolonged period of time. Dynorphin, though found in only minute quantities in nervous tissue, is important because it is an extremely powerful opiate, having 200 times as much pain-killing effect as morphine when injected directly into the analgesia system.

Thus, although all the fine details of the brain's opiate system are not yet entirely understood, nevertheless activation of the analgesia system either by nervous signals entering the periaqueductal gray area or by morphine-like drugs can totally or almost totally suppress many pain signals entering through the peripheral nerves

INHIBITION OF PAIN TRANSMISSION AT THE CORD LEVEL BY TACTILE SIGNALS

Another important landmark in the saga of pain control was the discovery that stimulation of large sensory fibers from the peripheral tactile receptors depresses the transmission of pain signals either from the same area of the body or even from areas sometimes located many segments away. This explains why such simple maneuvers as rubbing the skin near painful areas is often very effective in relieving pain. And it probably also explains why liniments are often useful in the relief of pain. This mechanism and simultaneous psychogenic excitation of the central analgesia system are probably also the basis of pain relief by acupuncture.

Treatment of Pain by Electrical Stimulation. Several clinical procedures have been developed recently for suppressing pain by stimulating large sensory nerve

fibers The stimulating electrodes are placed on selected areas of the skin, or on occasion they have been implanted over the spinal cord to stimulate the dorsal sensory columns. And, in a few patients, electrodes have even been placed stereotaxically in the periventricular or periaqueductal area of the diencephalon. The patient can then personally control the degree of stimulation. Dramatic relief has been reported in some instances

REFERRED PAIN

Often a person feels pain in a part of his or her body that is considerably removed from the tissues causing the pain This pain is called *referred pain*. Usually the pain is initiated in one of the visceral organs and referred to an area on the body surface Also, pain may be referred to another deep area of the body not exactly coincident with the location of the viscus producing the pain A knowledge of these different types of referred pain is extremely important in clinical diagnosis because many visceral ailments cause no other signs except referred pain

Mechanism of Referred Pain. Figure 50–7 illustrates the most likely mechanism by which most pain is referred In the figure, branches of visceral pain fibers are shown to synapse in the spinal cord with some of the same second order neurons that receive pain fibers from the skin When the visceral pain fibers are stimulated, pain signals from the viscera are then conducted through at least some of the same neurons that conduct pain signals from the skin, and the person has the feeling that the sensations actually originate in the skin itself

REFERRED PAIN CAUSED BY REFLEX MUSCULAR SPASM

Some types of referred pain are caused secondarily by reflex muscular spasm For instance, pain in a ureter can cause reflex spasm of the lumbar muscles Often the pain from the ureter itself is hardly felt at all, but instead almost all the pain results from spasm of the muscles

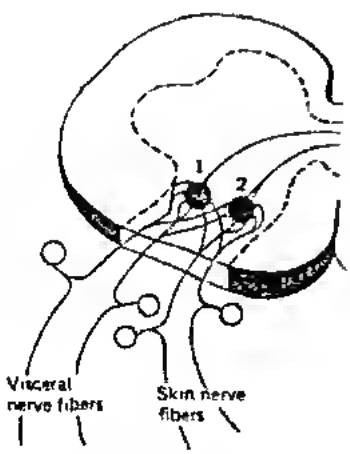


Figure 50–7. Mechanism of referred pain and referred hyperalgesia.

Many back pains and some types of headache also appear to be caused by muscular spasm, the spasm being excited reflexly from much weaker pain signals originating elsewhere in the body.

VISCERAL PAIN

In clinical diagnosis, pain from the different viscera of the abdomen and chest is one of the few criteria that can be used for diagnosing visceral inflammation, disease, and other ailments. In general, the viscera have sensory receptors for no other modalities of sensation besides pain Also, visceral pain differs from surface pain in several important aspects.

One of the most important differences between surface pain and visceral pain is that highly localized types of damage to the viscera rarely cause severe pain For instance, a surgeon can cut the gut entirely in two in a patient who is awake without causing significant pain. On the other hand, any stimulus that causes *diffuse stimulation of pain nerve endings* throughout a viscus causes pain that can be extremely severe. For instance, ischemia caused by occluding the blood supply to a large area of gut stimulates many diffuse pain fibers at the same time and can result in extreme pain

CAUSES OF TRUE VISCERAL PAIN

Any stimulus that excites pain nerve endings in diffuse areas of the viscera causes visceral pain. Such stimuli include ischemia of visceral tissue, chemical damage to the surfaces of the viscera, spasm of the smooth muscle in a hollow viscus, distention of a hollow viscus, or stretching of the ligaments.

Essentially all the true visceral pain originating in the thoracic and abdominal cavities is transmitted through sensory nerve fibers that run in the sympathetic nerves These fibers are small type C fibers and, therefore, can transmit only the chronic-burning-aching type of pain The pathways for transmitting true visceral pain will be discussed in more detail later in this chapter

Ischemia. Ischemia causes visceral pain in exactly the same way that it does in other tissues, presumably because of the formation of acidic metabolic end-products or tissue degenerative products, such as bradykinin, proteolytic enzymes, or others that stimulate the pain nerve endings

Chemical Stimuli. On occasion, damaging substances leak from the gastrointestinal tract into the peritoneal cavity For instance, proteolytic acidic gastric juice often leaks through a ruptured gastric or duodenal ulcer. This juice causes widespread digestion of the visceral peritoneum, thus stimulating extremely broad areas of pain fibers The pain is usually extremely severe.

Spasm of a Hollow Viscus. Spasm of the gut, the gallbladder, a bile duct, the ureter, or any other hollow viscus can cause pain in exactly the same way that spasm of skeletal muscle causes pain. This possibly results from mechanical stimulation of the pain endings Or its cause might be diminished blood flow to the muscle combined with increased metabolic need of the muscle for nutrients Thus, *relative* ischemia could develop, which causes severe pain.

Often, pain from a spastic viscus occurs in the form of *cramps*, the pain increasing to a high degree of

severity and then subsiding, this process continuing rhythmically once every few minutes. The rhythmic cycles result from rhythmic contraction of smooth muscle. For instance, each time a peristaltic wave travels along an overly excitable spastic gut, a cramp occurs. The cramping type of pain frequently occurs in gastroenteritis, constipation, menstruation, parturition, gallbladder disease, or ureteral obstruction.

Overdistension of a Hollow Viscus. Extreme overfilling of a hollow viscus also results in pain, presumably because of overstretch of the tissues themselves. However, overdistension can also collapse the blood vessels that encircle the viscus, or that pass into its wall, thus perhaps promoting ischemic pain.

Insensitive Viscera

A few visceral areas are almost entirely insensitive to pain of any type. These include the parenchyma of the liver and the alveoli of the lungs. Yet the liver capsule is extremely sensitive to both direct trauma and stretch, and the *bile ducts* are also sensitive to pain. In the lungs, even though the alveoli are insensitive, the bronchi and the parietal pleura are both very sensitive to pain.

"PARIETAL" PAIN CAUSED BY VISCERAL DAMAGE

In addition to true visceral pain, some pain sensations are also transmitted from the viscera through nonvisceral nerve fibers that innervate the parietal peritoneum, pleura, or pericardium.

Characteristics of Parietal Visceral Pain. When a disease affects a viscus, it often spreads to the parietal wall of the visceral cavity. This wall, like the skin, is supplied with extensive innervation from the spinal nerves, not from the sympathetic nerves, including the "fast" delta fibers which are different from the "slow" type C fibers in the true visceral pain pathways of the sympathetic nerves. Therefore, pain from the parietal wall of the visceral cavity is frequently very sharp and pricking in quality, though it can also have burning and aching qualities as well if the pain stimulus is diffuse. To emphasize the difference between this pain and true visceral pain: a knife incision through the *parietal* peritoneum is very painful, even though a similar cut through the visceral peritoneum or through a gut is not painful.

LOCALIZATION OF VISCERAL PAIN—THE "VISCERAL" AND THE "PARIETAL" TRANSMISSION PATHWAYS

Pain from the different viscera is frequently difficult to localize for a number of reasons. First, the brain does not know from firsthand experience that the different organs exist, and, therefore, any pain that originates internally can be localized only generally. Second, sensations from the abdomen and thorax are transmitted by two separate pathways to the central nervous system—the *true visceral pathway* and the *parietal pathway*. The true visceral pain is transmitted via sensory fibers of the autonomic nervous system, and the sensations are *referred* to surface areas of the body often far from the painful organ. On the other hand, parietal sensations are conducted *directly* into the local spinal nerves from the parietal peritoneum, pleura, or pericardium, and the sensations are usually *localized directly over the painful area*.

The True Visceral Pathway for Transmission of Pain. Most of the internal organs of the body are supplied by type C pain fibers that pass along the visceral sympathetic nerves into the spinal cord and thence up the lateral division of the anterolateral sensory pathway along with the pain fibers from the body's surface. A few visceral pain fibers—those from the distal portion of the colon, from the rectum, and from the bladder—enter the spinal cord through the sacral parasympathetic nerves, and some enter the central nervous system through various cranial nerves. These include fibers in the glossopharyngeal and vagus nerves which transmit pain from the pharynx, trachea, and upper esophagus. Some fibers from the surfaces of the diaphragm as well as from the lower esophagus are also carried in the phrenic nerves.

Localization of Referred Pain Transmitted by the Visceral Pathways. The position in the cord to which visceral afferent fibers pass from each organ, as well as the position on the surface of the body to which pain is referred—as illustrated in Figure 50–8, depends on the segment of the body from which the organ developed embryologically. For instance, the heart originated in the neck and upper thorax. Consequently, the heart's visceral pain fibers enter the cord all the way from segment C-3 down to T-5. The stomach had its origin approximately from the seventh to the ninth thoracic segments of the embryo, and consequently the visceral afferents from the stomach enter the spinal cord between these levels. The gallbladder had its origin almost entirely in the ninth thoracic segment, so that the visceral afferents from the gallbladder enter the spinal cord at T-9.

Because the visceral afferent pain fibers are responsible for transmitting referred pain from the viscera, the location of the referred pain on the surface of the body is in the dermatome of the segment from which the visceral organ was originally derived in the embryo.

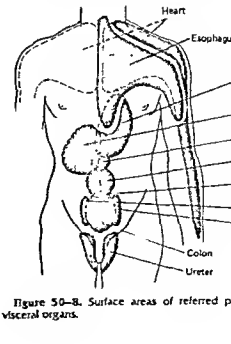


Figure 50–8. Surface areas of referred pain from different visceral organs.

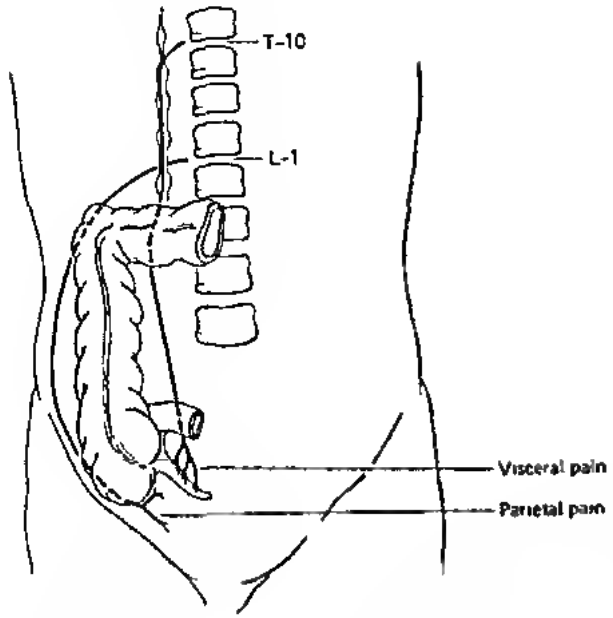


Figure 50–9. Visceral and parietal transmission of pain from the appendix.

Some of the areas of referred pain on the surface of the body are shown in Figure 50–8.

The Parietal Pathway for Transmission of Abdominal and Thoracic Pain. A second set of pain fibers penetrates inward from the spinal nerves that supply the external body surfaces to innervate as well the parietal peritoneum, parietal pleura, and parietal pericardium. Also, retroperitoneal visceral organs and perhaps portions of the mesentery are innervated to some extent by parietal pain fibers. The kidney, for instance, is supplied by both visceral and parietal fibers.

Pain from the viscera is frequently localized to two surface areas of the body at the same time because of the dual pathways for transmission of pain. Figure 50–9 illustrates dual transmission of pain from an inflamed appendix. Impulses pass from the appendix through the sympathetic visceral pain fibers into the sympathetic chain and then into the spinal cord at approximately T-10 or T-11; this pain is referred to an area around the umbilicus and is of the aching, cramping type. On the other hand, pain impulses also often originate in the parietal peritoneum where the inflamed appendix touches the abdominal wall, and these impulses pass through the spinal nerves into the spinal cord at a level of approximately L-1 or L-2. This pain is localized directly over the irritated peritoneum in the right lower quadrant of the abdomen and is of the sharp type.

VISCERAL PAIN FROM VARIOUS ORGANS

Cardiac Pain. Almost all pain that originates in the heart results from ischemia secondary to coronary sclerosis, which was discussed in Chapter 25. This pain is referred mainly to the base of the neck, over the shoulders, over the pectoral muscles, and down the arms. Most frequently, the referred pain is on the left side rather than on the right—probably because the left side of the heart is much more frequently involved in coronary disease than is the right side—but occasionally referred pain occurs on the right side of the body as well as on the left.

The pain impulses are conducted through sympathetic nerves passing to the middle cervical ganglia, to the stellate ganglia, and to the first four or five thoracic ganglia of the sympathetic chains. Then the impulses pass into the spinal cord through the second, third, fourth, and fifth thoracic spinal nerves.

Direct Parietal Pain from the Heart. When coronary ischemia is extremely severe, such as immediately after a coronary thrombosis, intense cardiac pain frequently occurs directly underneath the sternum simultaneously with pain referred to other areas. This direct pain from underneath the sternum is difficult to explain on the basis of the visceral nerve connections. Therefore, it is highly probable that sensory nerve endings passing from the heart through the pericardial reflections around the great vessels conduct this direct pain.

In addition to pain from the heart, other sensations may accompany coronary thrombosis. One of these is a tight, oppressive sensation beneath the sternum. The exact cause of this is unknown, but a possible cause is reflex spasm of blood vessels, bronchioles, or muscles in the chest region.

Esophageal Pain. Pain from the esophagus is usually referred to the pharynx, to the lower neck, to the arms, or to midline chest regions beginning at the upper portion of the sternum and ending approximately at the lower level of the heart. Irritation of the gastric end of the esophagus may cause pain directly over the heart, though the pain has nothing whatever to do with the heart. Such pain may be caused by spasm of the lower esophageal sphincter, or it may result from chemical, bacterial, or other types of inflammatory irritations.

Gastric Pain. Pain arising in the fundus of the stomach—usually caused by gastritis—is referred to the anterior surface of the chest or upper abdomen, from slightly below the heart to an inch or so below the xyphoid process. This pain is frequently characterized as burning pain, it, or pain from the lower esophagus, causes the condition known as "heartburn."

Most peptic ulcers occur within 1 to 2 inches on either side of the pylorus in the stomach or in the duodenum, and pain from such ulcers is usually referred to a surface point approximately midway between the umbilicus and the xiphoid process. The cause of ulcer pain is almost undoubtedly chemical, because when the acid juices of the stomach are not allowed to reach the pain fibers in the ulcer crater the pain does not exist. This pain is characteristically intensely burning.

Biliary and Gallbladder Pain. Pain from the bile ducts and gallbladder is localized in the midepigastrium in almost the same place as pain caused by peptic ulcers. Also, biliary and gallbladder pain is often burning, like that from ulcers, though cramps often occur too.

Biliary disease, in addition to causing pain on the abdominal surface, frequently refers pain to a small area on the back at the tip of the right scapula. This pain is transmitted through sympathetic afferent fibers that enter the ninth thoracic segment of the spinal cord.

Pancreatic Pain. Lesions of the pancreas, such as acute or chronic pancreatitis in which the pancreatic enzymes eat away the pancreas and surrounding structures, promote intense pain in areas both anterior to and behind the pancreas. It should be remembered that the pancreas is located beneath the parietal peritoneum and that it receives many parietal sensory fibers from the posterior abdominal wall. Therefore, the pain is usually localized directly behind the pancreas in the back and is severe and burning in character.

Renal Pain. The kidney, kidney pelvis, and ureters are all retroperitoneal structures and receive most of their pain fibers directly from skeletal nerves. Therefore, pain is usually felt directly behind the ailing structure. However, pain occasionally is referred via visceral afferents to the anterior abdominal wall below and about 2 inches to the side of the umbilicus.

Pain from the bladder is felt directly over the bladder, presumably because the bladder is well innervated by parietal pain fibers However, pain also is sometimes referred to the groin and testicles because some nerve fibers from the bladder apparently synapse in the cord in association with fibers from the genital areas

Uterine Pain. Both parietal and visceral afferent pain may be transmitted from the uterus. The low abdominal cramping pains of dysmenorrhea are mediated through the sympathetic afferents, and an operation to cut the hypogastric nerves between the hypogastric plexus and the uterus will in many instances relieve this pain On the other hand, lesions of the uterus that spread into the adnexa around the uterus, or lesions of the fallopian tubes and broad ligaments, usually cause pain in the lower back or side This pain is conducted over parietal nerve fibers and is usually sharper in nature than the diffuse cramping pain of true dysmenorrhea

SOME CLINICAL ABNORMALITIES OF PAIN AND OTHER SENSATIONS

HYPERALGESIA

A pain pathway may become excessively excitable, this gives rise to *hyperalgesia*, which means hypersensitivity to pain. The basic causes of hyperalgesia are (1) excessive sensitivity of the pain receptors themselves, which is called *primary hyperalgesia*, or facilitation of sensory transmission, which is called *secondary hyperalgesia*.

An example of primary hyperalgesia is the extreme sensitivity of sunburned skin. Secondary hyperalgesia frequently results from lesions in the spinal cord or in the thalamus Several of these will be discussed in subsequent sections.

THE THALAMIC SYNDROME

Occasionally the posterolateral branch of the posterior cerebral artery, a small artery supplying the posteroventral portion of the thalamus, becomes blocked by thrombosis so that the nuclei of this area of the thalamus degenerate, while the medial and anterior nuclei of the thalamus remain intact. The patient suffers a series of abnormalities, as follows: First, loss of almost all sensations from the opposite side of the body occurs because of destruction of the relay nuclei. Second, ataxia (inability to control movements precisely) may be evident because of loss of position and kinesthetic signals normally relayed through the thalamus to the cortex. Third, after a few weeks to a few months some sensory perception in the opposite side of the body returns, but strong stimuli are usually necessary to elicit this. When the sensations do occur, they are poorly—if at all—localized, almost always very painful, sometimes lancinating, regardless of the type of stimulus applied to the body. Fourth, the person is likely to perceive many affective sensations of extreme unpleasantness or, rarely, extreme pleasantness; the unpleasant ones are often associated with emotional tirades.

The medial nuclei of the thalamus are not destroyed by thrombosis of the artery. Therefore, it is believed that these nuclei become facilitated and give rise to the enhanced sensitivity to pain transmitted through the reticular system as well as to the affective perceptions

HERPES ZOSTER

Occasionally a herpes virus, which is believed to be identical to the chickenpox virus, infects a dorsal root ganglion. This causes severe pain in the dermatomal segment normally subserved by the ganglion, thus eliciting a segmental type of pain that circles halfway around the body The disease is called *herpes zoster*, and the pain is commonly called "shingles"

There are two possible causes of the pain of herpes zoster One is that the disease destroys mainly the large mechanoreceptor afferent sensory fibers This theoretically could reduce the normal inhibitory effect of these fibers on the pain pathway in the substantia gelatinosa of the dorsal horn. Therefore, pain signals could become exacerbated in the absence of this inhibition

A second possibility is that the irritated neuronal cell bodies of the root ganglion are stimulated to excessive activity and thereby cause the pain

TIC DOULOUREUX

Lancinating pains occur in some persons over one side of the face in part of the sensory distribution area of the fifth or ninth nerves; this phenomenon is called *tic douloureux* (or *trigeminal neuralgia* or *glossopharyngeal neuralgia*). The pains feel like sudden electric shocks, and they may appear for only a few seconds at a time or they may be almost continuous Often, they are set off by exceedingly sensitive "trigger areas" on the surface of the face, in the mouth or in the throat—almost always by a mechanoreceptive stimulus instead of a pain stimulus For instance, when the patient swallows a bolus of food, as the food touches a tonsil it might set off a severe lancinating pain in the mandibular portion of the fifth nerve

The pain of tic douloureux can usually be blocked by cutting the peripheral nerve from the hypersensitive area. The sensory portion of the fifth nerve is often sectioned immediately inside the cranium, where the motor and sensory roots of the fifth nerve can be separated so that the motor portions, which are needed for many of the jaw movements, are spared while the sensory elements are destroyed Obviously, this operation leaves the side of the face anesthetic, which in itself may be annoying. Furthermore, it is sometimes unsuccessful, indicating that the lesion might be more central than the nerves themselves.

THE BROWN-SÉQUARD SYNDROME

Obviously, if the spinal cord is transected entirely, all sensations and motor functions distal to the segments of transection are blocked, but if only one side of the spinal cord is transected, the so-called Brown-Séquard syndrome occurs. The following effects of such a transection occur, and these can be predicted from a knowledge of the cord fiber tracts illustrated in Figure 50–10: All motor functions are blocked on the side of the

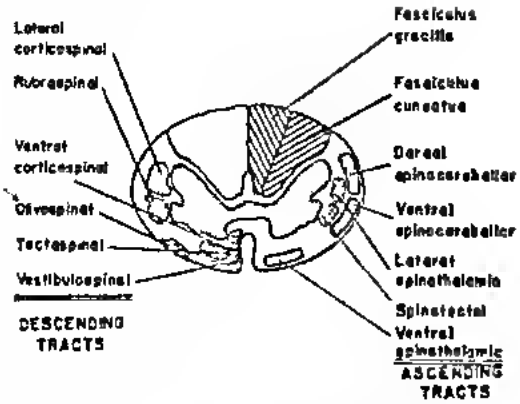


Figure 50–10. Cross-section of the spinal cord, showing principal ascending tracts on the right and principal descending tracts on the left

transection in all segments below the level of the transection, yet, only some of the modalities of sensation are lost on the transected side and others are lost on the opposite side. The sensations of pain, heat, and cold are lost on the opposite side of the body in all dermatomes two to six segments below the level of the transection. The sensations that are transmitted only in the dorsal and dorsolateral columns—kinesthetic and position sensations, vibration sensation, discrete localization, and two-point discrimination—are lost entirely on the side of the transection in all dermatomes below the level of the transection. Touch is impaired on the side of the transection because the principal pathway for transmission of light touch, the dorsal columns, is transected. Yet, "crude touch," which is poorly localized, still persists because of transmission in the opposite ventral spinothalamic tract.

HEADACHE

Headaches are actually referred pain to the surface of the head from the deep structures. Many headaches result from pain stimuli arising inside the cranium, but equally as many probably result from pain arising outside the cranium.

HEADACHE OF INTRACRANIAL ORIGIN

Pain-Sensitive Areas in the Cranial Vault. The brain itself is almost totally insensitive to pain. Even cutting or electrically stimulating the somatic sensory areas of the cortex only occasionally causes pain, instead, it causes tactile pins and needles types of paresthesias on the area of the body represented by the portion of the sensory cortex stimulated. Therefore, it is obvious from the outset that much or most of the pain of headache probably is not caused by damage within the brain itself.

On the other hand, *tugging on the venous sinuses, damaging the tentorium,* or *stretching the dura at the base of the brain* can all cause intense pain that is recognized as headache. Also, almost any type of traumatizing, crushing, or stretching stimulus to the *blood vessels of the dura* can cause headache. A very sensitive structure is the middle meningeal artery, and neurosurgeons are careful to anesthetize this artery specifically when performing brain operations under local anesthesia.

Areas of the Head to Which Intracranial Headache Is Referred. Stimulation of pain receptors in the intracranial vault above the tentorium, including the upper surface of the tentorium itself, initiates impulses in the fifth nerve and, therefore, causes referred headache to the front half of the head in the area supplied by the fifth cranial nerve, as illustrated in Figure 50–11.

On the other hand, pain impulses from beneath the tentorium enter the central nervous system mainly through the second cervical nerve, which also supplies the scalp behind the ear. Therefore, subtentorial pain stimuli cause "occipital headache" referred to the posterior part of the head as shown in Figure 50–11.

Types of Intracranial Headache. ***Headache of Meningitis.*** One of the most severe headaches of all is that resulting from meningitis, which causes inflammation of all the meninges, including the sensitive areas of the dura and the sensitive areas around the venous sinuses. Such intense damage as this can cause extreme headache pain referred over the entire head.

Headache Resulting from Direct Meningeal Trauma. Following a brain operation the patient ordinarily has intense headache for several days to several weeks. Though part of this headache may result from the brain trauma, experiments indicate that most of it results from meningeal irritation.

Another type of meningeal trauma that almost invariably causes headache is the meningeal irritation resulting from brain tumor. Usually, tumor headache is referred to a localized area of the head, the exact area depending on the portion of the meninges affected by the tumor. Since a tumor above the tentorium is likely to refer its pain to the front half of the head and any tumor below the tentorium to the occipital region of the skull, the general location of an intracranial tumor can often be predicted from the area of the headache.

Headache Caused by Low Cerebrospinal Fluid pressure. Removing as little as 20 ml of fluid from the spinal canal, particularly if the person remains in the upright position, often causes intense intracranial headache. Removing this quantity of fluid removes the flotation for the brain that is normally provided by the

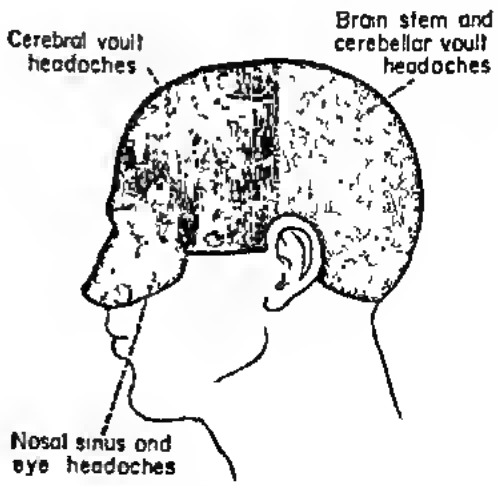


Figure 50–11. Areas of headache resulting from different causes.

cerebrospinal fluid. Therefore, the weight of the brain stretches the various dural surfaces and thereby elicits the pain which causes the headache.

Migraine Headache. Migraine headache is a special type of headache that is thought to result from abnormal vascular phenomena, though the exact mechanism is unknown.

Migraine headaches often begin with various prodromal sensations, such as nausea, loss of vision in part of the field of vision, visual aura, or other types of sensory hallucinations. Ordinarily, the prodromal symptoms begin half an hour to an hour prior to the beginning of the headache itself. Therefore, any theory that explains migraine headache must also explain these prodromal symptoms.

One of the *theories* of the cause of migraine headaches is that prolonged emotion or tension causes reflex vasospasm of some of the arteries of the head, including arteries that supply the brain itself The vasospasm theoretically produces ischemia of portions of the brain, and this is responsible for the prodromal symptoms. Then, as a result of the intense ischemia, something happens to the vascular wall to allow it to become flaccid and incapable of maintaining vascular tone for 24 to 48 hours. The blood pressure in the vessels causes them to dilate and pulsate intensely, and it is postulated that the excessive stretching of the walls of the arteries—including the extracranial arteries such as the temporal artery—causes the actual pain of migraine headaches However, it is possible that diffuse aftereffects of ischemia in the brain itself are at least partially if not mainly responsible for this type of headache.

Alcoholic Headache. As many people have experienced, a headache usually follows an alcoholic binge. It is most likely that alcohol, because it is toxic to tissues, directly irritates the meninges and causes intracranial pain.

Headache Caused by Constipation. Constipation causes headache in many persons Because it has been shown that constipation headache can occur in persons whose spinal cords have been cut, we know that this headache is not caused by nervous impulses from the colon. Therefore, it possibly results from absorbed toxic products or from changes in the circulatory system resulting from loss of fluid into the gut.

EXTRACRANIAL TYPES OF HEADACHE

Headache Resulting from Muscular Spasm. Emotional tension often causes many of the muscles of the head, including especially those muscles attached to the scalp and the neck muscles attached to the occiput, to become spastic, and it is postulated that this is one of the common causes of headache. The pain of the spastic head muscles supposedly is referred to the overlying areas of the head and gives one the same type of headache as do intracranial lesions.

Headache Caused by Irritation of the Nasal and Accessory Nasal Structures. The mucous membranes of the nose and also of all the nasal sinuses are sensitive to pain, but not intensely so. Nevertheless, infection or other irritative processes in widespread areas of the nasal structures usually cause headache that is referred behind the eyes or, in the case of frontal sinus infection, to the frontal surfaces of the forehead and scalp, as illustrated in Figure 50–11. Also, pain from the lower sinuses—such as the maxillary sinuses—can be felt in the face.

Headache Caused by Eye Disorders. Difficulty in focusing one's eyes clearly may cause excessive contraction of the ciliary muscles in an attempt to gain clear vision. Even though these muscles are extremely small, tonic contraction of them can be the cause of retroorbital headache. Also, excessive attempts to focus the eyes can result in reflex spasm in various facial and extraocular muscles, which is also a possible cause of headache.

A second type of headache originating in the eyes occurs when the eyes are exposed to excessive irradiation by ultraviolet light rays. Watching the sun or the arc of an arc-welder for even a few seconds may result in headache that lasts from 24 to 48 hours. The headache sometimes results from "actinic" irritation of the conjunctivae, and the pain is referred to the surface of the head or retro-orbitally However, focusing intense light from an arc or the sun on the retina can actually burn the retina, and this could result in headache.

THERMAL SENSATIONS

THERMAL RECEPTORS AND THEIR EXCITATION

The human being can perceive different gradations of cold and heat, progressing from *freezing cold* to *cold* to *cool* to *indifferent* to *warm* to *hot* to *burning hot*.

Thermal gradations are discriminated by at least three different types of sensory receptors: the cold receptors, the warmth receptors, and pain receptors. The pain receptors are stimulated only by extreme degrees of heat or cold and therefore are responsible, along with the cold and warmth receptors, for "freezing cold" and "burning hot" sensations.

The cold and warmth receptors are located immediately under the skin at discrete but separated points, each having a stimulatory diameter of about 1 mm In most areas of the body there are three to ten times as many cold receptors as warmth receptors, and the number in different areas of the body varies from as great as 15 to 25 cold points per square centimeter in the lips, to 3 to 5 cold points per square centimeter in the finger, to less than 1 cold point per square centimeter in some broad surface areas of the trunk. There are correspondingly fewer numbers of warmth points.

Though it is quite certain on the basis of psychological tests that there are distinctive warmth nerve endings, these have not yet been identified histologically. They are presumed to be free nerve endings because warmth signals are transmitted mainly over type C nerve fibers at transmission velocities of only 0.4 to 2 meters per second.

On the other hand, a definitive cold receptor has been identified. It is a special, small, type Aδ myelinated nerve ending that branches a number of times, the tips of which protrude into the bottom surfaces of basal epidermal cells. Signals are transmitted from these receptors via delta nerve fibers at velocities of up to about 20 meters per second. However, some cold sensations are also transmit-

ted in type C nerve fibers, which suggests that some free nerve endings also might function as cold receptors.

Stimulation of Thermal Receptors—Sensations of Cold, Cool, Indifferent, Warm, and Hot. Figure 50–12 illustrates the effects of different temperatures on the responses of four different nerve fibers: (1) a pain fiber stimulated by cold, (2) a cold fiber, (3) a warmth fiber, and (4) a pain fiber stimulated by heat. Note especially that these fibers respond differently at different levels of temperature. For instance, in the very cold region only the pain fibers are stimulated (if the skin becomes even colder so that it nearly freezes or actually does freeze, even these fibers cannot be stimulated). As the temperature rises to 10 to 15° C, pain impulses cease, but the cold receptors begin to be stimulated. Then, above about 30° C the warmth receptors become stimulated while the cold receptors fade out at about 43° C. Finally, at around 45° C, pain fibers begin to be stimulated by heat.

One can understand from Figure 50–12, therefore, that a person determines the different gradations of thermal sensations by the relative degrees of stimulation of the different types of endings. For instance, at 20° C only cold endings are stimulated, whereas at 44° C only warm endings are stimulated, at 35° C both cold and warm endings are stimulated; at 60° C pain endings are again stimulated. One can understand also from this figure why extreme degrees of cold or heat can both be painful and why both these sensations, when intense enough, may give almost exactly the same quality of sensations—that is, freezing cold and burning hot sensations feel almost alike; they are both very painful. (Also, not shown in the figure is the fact that high temperatures, as well as cold temperatures, excite the cold endings, which also makes hot temperatures often feel cold and painful.)

Stimulatory Effects of Rising and Falling Temperature—Adoption of Thermal Receptors. When a cold receptor is suddenly subjected

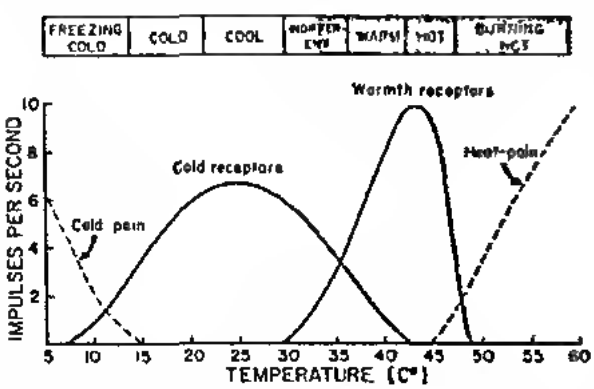


Figure 50–12. Frequencies of discharge of (1) a cold-pain fiber, (2) a cold fiber, (3) a warmth fiber, and (4) a heat pain fiber. (The responses of these fibers are drawn from original data collected in separate experiments by Zotterman, Hensel, and Kenshalo.)

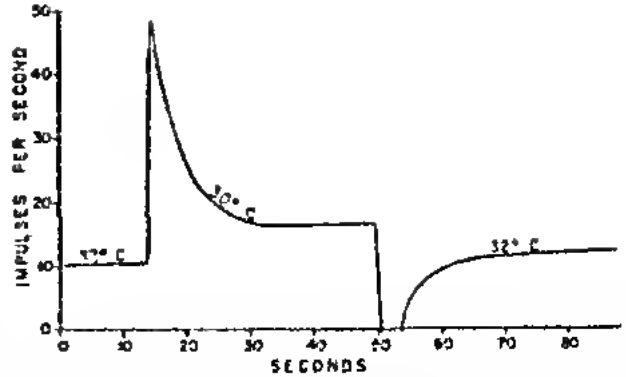


Figure 50–13. Response of a nerve fiber from a cold receptor following, first, instantaneous change in skin temperature from 32° to 30° C and, second, instantaneous change back to 32° C. Note the adaptation of the receptor and also the higher steady state level of discharge at 30° than at 32°.

to an abrupt fall in temperature, it becomes strongly stimulated at first, but this stimulation fades rapidly during the first few seconds and progressively more slowly during the next half hour or more. In other words, the receptor adapts to a very great extent; this is illustrated in Figure 50–13, which shows that the frequency of discharge of a cold receptor rose approximately fourfold when the temperature fell suddenly from 32° to 30° C, but in less than a minute the frequency fell about five sixths of the way back to the original control value. Later, the temperature was suddenly raised from 30° to 32° C. The cold receptor stopped firing entirely for a short time but after adapting returned to its original control level.

Thus, it is evident that the thermal senses respond markedly to *changes in temperature* in addition to being able to respond to steady states of temperature, as was depicted in Figure 50–12. This means, therefore, that when the temperature of the skin is actively falling, a person feels much colder than when the temperature remains at the same level. Conversely, if the temperature is actively rising the person feels much warmer than he or she would at the same temperature if it were constant.

The response to changes in temperature explains the extreme degree of heat that one feels on first entering a tub of hot water and the extreme degree of cold felt on going from a heated room to the out-of-doors on a cold day.

Mechanism of Stimulation of the Thermal Receptors. It is believed that the thermal receptors are stimulated by changes in their metabolic rates, these changes resulting from the fact that temperature alters the rates of intracellular chemical reactions more than twofold for each 10° C change. In other words, thermal detection probably results not from direct physical stimulation but instead from chemical stimulation of the endings as modified by the temperature.

Spatial Summation of Thermal Sensations. Because the number of cold or warmth endings in any one surface area of the body is very slight, it

is difficult to judge gradations of temperature when small areas are stimulated. However, when a large area of the body is stimulated all at once, the thermal signals from the entire area summate. Indeed, one reaches the maximum ability to discern minute temperature variations when the entire body is subjected to a temperature change all at once. For instance, rapid changes in temperature of as little as 0.01° C can be detected if this change affects the entire surface of the body simultaneously. On the other hand, temperature changes 100 times this great might not be detected when the skin surface affected is only a square centimeter or so in size.

TRANSMISSION OF THERMAL SIGNALS IN THE NERVOUS SYSTEM

In general, thermal signals are transmitted in almost parallel, but not the same, pathways as pain signals. On entering the spinal cord, the signals travel for a few segments upward or downward in the *tract of Lissauer* and then terminate mainly in laminae I, II, and III of the dorsal horns. After a small amount of processing by one or more cord neurons, the signals enter long, ascending thermal fibers that cross to the lateral division of the opposite anterolateral sensory tract and terminate in (a) the reticular areas of the brain stem and (b) the ventrobasal complex of the thalamus. A few thermal signals are also relayed to the somatic sensory cortex from the ventrobasal complex. Occasionally, a neuron in somatic sensory area I has been found by microelectrode studies to be directly responsive to either cold or warm stimuli in specific areas of the skin. Furthermore, it is known that removal of the postcentral gyrus in the human being reduces the ability to distinguish gradations of temperature.

REFERENCES

Akil, H, *et al* Endogenous opioids Etiology and function *Annu Rev Neurosci* 7 223, 1984

American Physiological Society Sensory Processes Washington, D C, American Physiological Society, 1984

Basbaum, A I, and Fields H L. Endogenous pain control systems Brainstem spinal pathways and endorphin circuitry *Annu Rev Neurosci*, 7 303, 1984

Berger, P. A, *et al* Behavioral pharmacology of the endorphins *Annu Rev Med*, 33 397, 1982

Bond, M R Pain—Its Nature, Analysis, and Treatment. New York, Churchill Livingstone, 1984

Brandes, J B Control of Migraine New York, W W Norton & Co, 1979,

Coren, S, *et al* Sensation and Perception New York, Academic Press, 1979

Currie, D J. Abdominal Pain Washington, D C, Hemisphere Publishing Corporation, 1979

Darian Smith, I, *et al* Posterior parietal cortex Relations of unit activity to sensorimotor function *Annu Rev Physiol*, 47 141, 1979

Darian-Smith, I • The sense of touch Performance and peripheral neural process In Darian-Smith, I. (ed) Handbook of Physiology Sec 1, Vol III Bethesda, American Physiological Society, 1984, p 739

Darian Smith, I • Thermal sensibility *In* Darian-Smith, I. (ed) Handbook of Physiology Sec. 1, Vol III Bethesda, American Physiological Society, 1984, p 879

Dubner, R, and Bennett, G J Spinal and trigeminal mechanisms of nociception *Annu Rev. Neurosci*, 6 381, 1983

Fields, H L., and Dasbaum, A. I Brainstem control of spinal pain-transmission neurons *Annu Rev. Physiol*, 40 217, 1978

Friedhoff, A J, and Miller, J. C: Clinical implications of receptor sensitivity modification. *Annu Rev Neurosci*, 6 121, 1983

Gelmers, H J, Calcium-channel blockers in the treatment of migraine *Am. J Cardiol*, 55 1398, 1985

Goldstein, E B• Sensation and Perception Belmont, Cal, Wadsworth Publishing Co, 1980

Graham, M. (ed): Disorders of the Facial Nerve, Anatomy, Diagnosis and Management. New York, Raven Press, 1982

Guyton, A. C, and Reeder, R. C • Pain and contracture in poliomyelitis. *Arch Neurol Psychiatr*, 63 954, 1950

Haft, J I (ed) Differential Diagnosis of Chest Pain and Other Cardiac Symptoms *Mt Kisco, N Y, Futura Publishing Co, 1983*,

Han, J. S, and Terenius, L. Neurochemical basis of acupuncture analgesia *Annu Rev. Pharmacol Toxicol*, 22 193, 1982

Hardy, J D The nature of pain *J Chronic Dis*, 4 22, 1956

Hardy, J D, *et al* Pain Sensations and Reactions. Baltimore, Williams & Wilkins, 1952

Health and Public Policy Committee, American College of Physicians Biofeedback for headaches *Ann. Intern Med*, 102 128, 1985

Hochberg, J Perception *In* Darian-Smith, I (ed) Handbook of Physiology Sec 1, Vol III Bethesda, American Physiological Society, 1984, p 75

Hyvarinen, J Posterior parietal lobe of the primate brain *Physiol Rev*, 62 1060, 1982

Ichia, S (ed) Pain Treatment Pituitary Neuroandenolysis in the Treatment of Cancer Pain and Hormone Dependent Tumors New York, Raven Press, 1983

Jacob, J (ed) Receptors New York, Pergamon Press, 1979,

Jung R. Sensory research in historical perspective Some philosophical foundations of perception *In* Darian Smith, I. (ed) Handbook of Physiology Sec 1, Vol III Bethesda, American Physiological Society, 1984, p 1

Kaas, J H What, if anything, is SI? Organization of first somatosensory area of cortex *Physiol Rev*, 63 206, 1983

Kaas, J, *et al* The reorganization of the somatosensory cortex following peripheral nerve damage in adult and developing mammals. *Annu Rev Neurosci*, 6 325, 1983

Kenshalo, D R (ed) Sensory Function of the Skin of Humans New York, Plenum Press, 1979

Kruger, L. (ed) Neural Mechanisms of Pain New York, Raven Press, 1984

Lewis, R V, and Stern, A S Biosynthesis of the enkephalins and enkephalin-containing polypeptides *Annu Rev Pharmacol Toxicol*, 22 353, 1983

Mountcastle, V B Central nervous mechanisms in mechanoreceptive sensibility *In* Darian-Smith, I (ed) Handbook of Physiology. Sec. 1, Vol III Bethesda, American Physiological Society, 1984, p. 789

Neff, W D (ed) Contributions to Sensory Physiology New York, Academic Press, 1982

Norrsell, U Behavioral studies of the *somatosensory system Physiol Rev*, 60 327, 1980

Olton D S Spatial memory *Sci Am*, 236(6) 82 1977

Perl, E R. Pain and nociception *In* Darian Smith, I (ed) Handbook of Physiology Sec. 1, Vol III Bethesda, American Physiological Society, 1984, p 915

Perl, E R, and Borvie, J G Neural substrates of somatic sensation. *In* MIP International Review of Science Physiology Vol. 3 Baltimore, University Park Press, 1974, p 303

Ray, Practical Management of Pain Chicago, Year Book Medical Publishers, 1985

Raskin, N H Chemical headaches *Annu Rev Med*, 32 63, 1981.

Raskin, N H, and Appenzeller, D Headache Philadelphia, W. B Saunders Co, 1980

Schiebel, A B The brain stem reticular core and sensory function *In* Darian-Smith, I. (ed) Handbook of Physiology Sec. 1, Vol III. Bethesda, American Physiological Society, 1984, p 213

Seltzer, S. Pain in Dentistry Diagnosis and Management. Philadelphia, J B Lippincott Co, 1978

Silen, W (rev) Cope's Early Diagnosis of the Acute Abdomen New York, Oxford University Press, 1979

Stebbins, W C, Brown, C. H, and Petersen, M R. Sensory function in animals *In* Darian Smith, I (ed.) Handbook of Physiology. Sec. 1, Vol III Bethesda, American Physiological Society, 1984, p 123

Stumsel, B • Pain, Analgesia, and Addiction The Pharmacologic Treatment of Pain New York, Raven Press, 1983

Weiss, T. F Relation of receptor potentials of cochlear hair cells to spike discharges of cochlear neurons. *Annu Rev. Physiol*, 46 247, 1984

Woolsey, C. N. (ed.) Cortical Sensory Organization Multiple Somatic Areas. Clifton, N.J, Humana Press, 1981.

Zimmerman, M• Neurophysiology of nociception *Int. Rev Physiol*, 10 179, 1976

51

Motor Functions of the Spinal Cord; and the Cord Reflexes

In the discussion of the nervous system thus far, we have considered principally the input of sensory information. In the following chapters we will discuss the origin and output of motor signals, the signals that cause muscle contraction and other motor effects throughout the body. Sensory information is integrated at all levels of the nervous system and causes appropriate motor responses, beginning in the spinal cord with relatively simple reflexes, extending into the brain stem with still more complicated responses, and, finally, extending to the cerebrum where the most complicated responses are controlled. The present chapter discusses the control of motor function especially in response to spinal cord reflexes.

Experimental Preparations for Studying Cord Reflexes—The Spinal Animal and the Decerebrate Animal. Normally, most functions of the spinal cord are strongly influenced by signals from the brain. Therefore, to study the isolated cord reflexes it is necessary to separate the cord from the higher centers. This is usually done in one of two different types of preparations. (1) the *spinal animal* or (2) the *decerebrate animal.*

The spinal animal is prepared by cutting the spinal cord at any level above the region in which the cord reflexes are to be studied. For a variable period of time, depending upon the phylogenetic level of the animal, as will be discussed later in the chapter, the cord reflexes are deeply depressed immediately after removal of the signals from the brain. However, over a period of minutes, hours, or days in animals and over a period of weeks in the human being, the cord reflexes become progressively more active and can then be studied independently of control by the upper levels of the nervous system. Often the cord reflexes become even more excitable under these conditions than normally, as will be discussed later in the chapter.

In the decerebrate preparation, the brain stem is usually transected between the superior and inferior colliculi. Transecting the brain stem at this level removes the voluntary control centers of the forebrain, and it also removes inhibitory influences from the basal ganglia that normally suppress the activities of the lower brain stem and spinal cord; this will be discussed in detail in the following chapter. Removal of this inhibition causes an immediate increase in the activity of many of the cord reflexes, especially the antigravity reflexes that help the animal hold itself up against gravity. This preparation has an advantage over the spinal animal in that many of the cord reflexes are exacerbated even in the acute preparation and, therefore, can be studied immediately, though, of course, lower brain stem reflexes may still interfere with the cord functions.

ORGANIZATION OF THE SPINAL CORD FOR MOTOR FUNCTIONS

The cord gray matter is the integrative area for the cord reflexes and other motor functions. Figure 51-1 shows the typical organization of the cord gray matter in a single cord segment. Sensory signals enter the cord through the sensory roots. After entering the cord, every sensory signal travels to two separate destinations. First, the sensory nerve or its collaterals terminate in the gray matter of the cord and elicit local segmental responses—local excitatory effects, facilitatory effects, reflexes, and so forth. Second, the signals

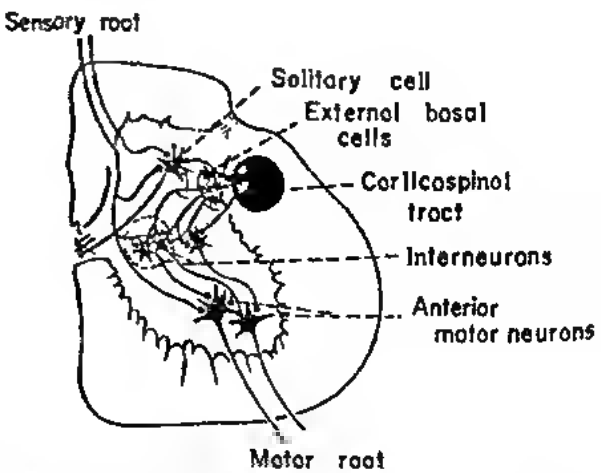


Figure 51-1. Connections of the sensory fibers and corticospinal fibers with the interneurons and anterior motor neurons of the spinal cord

travel to higher levels of the nervous system—to higher levels in the cord itself, to the brain stem, or even to the cerebral cortex, as described in the past few chapters.

Each segment of the spinal cord has several million neurons in its gray matter. Aside from the sensory relay neurons discussed in Chapters 49 and 50, the remainder of these neurons are of two separate types, the *anterior motor neurons* and the *interneurons*.

The Anterior Motor Neurons. Located in each segment of the anterior horns of the cord gray matter are several thousand neurons that are 50 to 100 per cent larger than most of the others and are called *anterior motor neurons*. These give rise to the nerve fibers that leave the cord via the anterior roots and innervate the skeletal muscle fibers. The neurons are of two types, the *alpha motor neurons* and the *gamma motor neurons*

The Alpha Motor Neurons. The alpha motor neurons give rise to large, type A alpha ($A\alpha$) nerve fibers ranging from 9 to 20 microns in diameter that innervate the large skeletal muscle fibers. Stimulation of a single nerve fiber excites from as few as three to as many as several hundred skeletal muscle fibers, which are collectively called the *motor unit*. Transmission of nerve impulses into skeletal muscles and their stimulation of the muscle fibers were discussed in Chapters 10 through 12.

The Gamma Motor Neurons. In addition to the alpha motor neurons that excite contraction of the skeletal muscle fibers, about one half as many much smaller gamma motor neurons are located along with the alpha motor neurons in the anterior horns. These transmit impulses through type A gamma ($A\gamma$) fibers, averaging 5 microns in diameter, to very small, special skeletal muscle fibers called *intrafusal fibers*. These are part of the *muscle spindle*, which is discussed later in the chapter.

The Interneurons. The interneurons are present in all areas of the cord gray matter—in the dorsal horns, in the anterior horns, and in the intermediate areas between these two. These cells are numerous—approximately 30 times as numerous as the anterior motor neurons. They are small and highly excitable, often exhibiting spontaneous activity and capable of firing as rapidly as 1500 times per second. They have many interconnections one with the other, and many of them directly innervate the anterior motor neurons as illustrated in Figure 51–1. The interconnections among the interneurons and anterior motor neurons are responsible for many of the integrative functions of the spinal cord that are discussed in the remainder of this chapter.

Essentially all the different types of neuronal circuits described in Chapter 47 are found in the interneuron pool of cells of the spinal cord, including the *diverging*, *converging*, and *repetitive-discharge* circuits. In this chapter we will see many applications of these different circuits to the performance of specific reflex acts by the spinal cord.

Only a few incoming sensory signals from the spinal nerves or signals from the brain terminate directly on the anterior motor neurons. Instead, most of them are transmitted first through interneurons where they are appropriately processed. Thus, in Figure 51–1, it is shown that the corticospinal tract terminates almost entirely on interneurons, and it is only after the signals from this tract have been integrated in the interneuron pool with signals from other spinal tracts or from the spinal nerves that they finally impinge on the anterior motor neurons to control muscular function.

The Renshaw Cell Inhibitory System. Located also in the ventral horns of the spinal cord in close association with the motor neurons are a large number of small interneurons called *Renshaw cells*. Almost immediately after the motor axon leaves the motor neuron, collateral branches from the axon pass to the adjacent Renshaw cells. These in turn are inhibitory cells that transmit inhibitory signals to the nearby motor neurons. Thus, stimulation of each motor neuron, tends to inhibit the surrounding motor neurons, an effect called *recurrent inhibition*. This effect is probably important for the following major reason:

It shows that the motor system utilizes the principle of lateral inhibition to focus, or sharpen, its signals in the same way that the sensory system utilizes this principle—that is, to allow unabated transmission of the primary signal while suppressing the tendency for signals to spread to adjacent neurons

Multisegmental Connections in the Spinal Cord—The Propriospinal Fibers

More than half of all the nerve fibers ascending and descending in the spinal cord are *propriospinal fibers* These are fibers that run from one segment of the cord to another. In addition, the terminal fibrils of sensory fibers as they enter the cord branch both up and down the spinal cord, some of the branches transmitting signals only a segment or two in each direction whereas others transmit signals many segments. These ascending and descending fibers of the cord provide pathways for the multisegmental reflexes that will be described later in this chapter, including reflexes that coordinate movements in the forelimbs and hindlimbs simultaneously.

THE MUSCLE RECEPTORS—MUSCLE SPINDLES AND GOLGI TENDON ORGANS—AND THEIR ROLES IN MUSCLE CONTROL

Proper control of muscle function requires not only excitation of the muscle by the anterior motor neurons but also continuous feedback of information from each muscle to the nervous system, giving the status of the muscle at each instant. That is, what is the length of the muscle, what is

its instantaneous tension, and how rapidly is its length or tension changing? To provide this information, the muscles and their tendons are supplied abundantly with two special types of sensory receptors: (1) *muscle spindles*, which are distributed throughout the belly of the muscle and which send information to the nervous system about either the muscle length or rate of change of its length, and (2) *Golgi tendon organs*, which transmit information about tension or rate of change af tension.

The signals from these two receptors are either entirely or almost entirely for the purpose of muscle control itself, because they operate almost entirely at a subconscious level. Even so, they transmit tremendous amounts of information into the spinal cord, the cerebellum, and even the cerebral cortex, helping each of these portions of the nervous system in its function for controlling muscle contraction

RECEPTOR FUNCTION OF THE MUSCLE SPINDLE

Structure and Innervatlan of the Musele Spindle. The physiologic arganization of the muscle spindle is illustrated in Figure 51–2 Each spindle is built around 3 to 12 small *intrafusol muscle fibers* that are pointed at their ends and are attached to the glycocalyx of the surrounding *extrafusol* skeletal muscle fibers. Each intrafusal fiber is a very small skeletal muscle fiber However, the central region of each of these fibers—that is, the area midway between its two ends—has either no or few actin and myosin filaments Therefore, this central portion does not contract when the ends do Instead, it functions as a sensory receptor as we shall describe later The end portions are excited by the small *gamma motor nerve fibers* originating from the gamma motor neurons that were described earlier These fibers are frequently also called *gamma efferent fibers*, in contradistinction to the *alpha efferent fibers* that innervate the extrafusal skeletal muscle

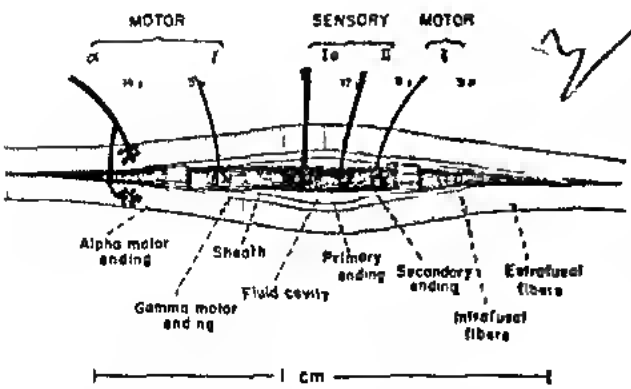


Figure 51–2. The muscle spindle, showing its relationship to the large extrafusal skeletal muscle fibers Note also both the motor and the sensory innervation of the muscle spindle

Excitation of the Spindle Receptors

The receptor portion of the muscle spindle is its central portion where the intrafusal muscle fibers have no contractile elements. As illustrated in Figure 51–2, and also in more detail in Figure 51–3, sensory fibers originate in this area, and these are stimulated by stretch of this midportion of the spindle One can readily see that the muscle spindle receptor can be excited in two different ways.

(1) Lengthening the whole muscle will obviously stretch the midportion of the spindle and therefore excite the receptor

(2) Even if the length of the entire muscle does not change, contraction of the end-portions of the intrafusal fibers will also stretch the midportions of the spindle and therefore excite the receptor.

Twa types of sensory endings are found in the receptor area of the muscle spindle. These are:

The Primary Ending. In the very center of the receptor area a large sensory fiber encircles the central portion of each intrafusal fiber, forming the so-called *primary ending* or also called the *annulospiral ending* This nerve fiber is a type Ia fiber averaging 17 microns in diameter, and it transmits sensory signals to the spinal cord at a velocity of 70 to 120 meters per secand, as rapidly as any type of sensory nerve fiber in the entire body

The Secondary Endings. Usually one, but sometimes two, smaller sensory nerve fiber, a type II fiber with an average diameter of 8 microns, innervates the receptor region on ane side of the primary ending, as illustrated in Figures 51–2 and 51–3. This sensory ending is called the *secondary ending*, or sometimes the *flower spray ending* because in some preparations it looks like a flower spray even though it mainly encircles the intrafusal fibers in the same way that the type Ia fiber does

Division of the Intrafusal Fibers into Nuclear Bag and Nuclenr Chaln Fibers—Dynamic nnd Sttatic Responses of the Muscle Spindle. There are also two different types of intrafusal fibers: (1) *nuclear bag fibers* (one to three in each spindle), in which a large number of nuclei are congregated into a bag in the central portion of the receptor area, and (2) *nuclear chain fibers* (three to nine), which are about half as large in diameter and half as long as the nuclear bag fibers and have nuclei aligned in a chain thraughout the receptor area. The primary ending innervates both the nuclear bag intrafusal fibers *and* the nuclear chain fibers. On the other hand, the secondary ending usually innervates only the nuclear chain fibers. These relationships are all illustrated in Figure 51–3.

Response of Both the Primary and Secondary Endings to Length of the Receptor—The "Static" Response. When the receptor portion of the muscle spindle is stretched slowly, the number

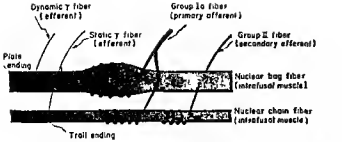


Figure 51–3. Details of nerve connections to the nuclear bag and nuclear chain muscle spindle fibers (Modified from Stein *Physiol Rev*, 54 225, 1974, and Boyd *Philos Trans R Soc Lond [Biol Sci]*, 245 81, 1962)

of impulses transmitted from both the primary and the secondary endings increases almost directly in proportion to the degree of stretch, and the endings continue to transmit these impulses for many minutes. This effect is called the *static response* of the spindle receptor, meaning simply that both the primary and the secondary endings continue to transmit their signals for as long as the receptor itself remains stretched. Because only the *nuclear chain* type of intrafusal fiber is innervated by both the primary and the secondary endings, it is believed that these nuclear chain fibers are responsible for the static response.

Response of the Primary Ending (But Not the Secondary Ending) to Rate of Change of Receptor Length—The "Dynamic" Response. When the length of the spindle receptor increases suddenly, the primary ending (but not the secondary ending) is stimulated especially powerfully, much more powerfully than the stimulus caused by the static response. This excess stimulus of the primary ending is called the *dynamic response*, which means that the primary ending responds extremely actively to a rapid *rate of change* in length. When the length of a spindle receptor increases only a fraction of a micron, if this increase occurs in a fraction of a second the primary receptor transmits tremendous numbers of excess impulses into the Ia fiber, but only *while the length is actually increasing*. As soon as the length has stopped increasing, the rate of impulse discharge returns almost back to its original level, except for the small static response that is still present in the signal.

Conversely, when the spindle receptor shortens, this change momentarily decreases the impulse output from the primary ending; then, as soon as the receptor area has reached its new shortened length, the impulses reappear in the Ia fiber within a fraction of a second. Thus, the primary ending sends extremely strong signals to the central nervous system to apprise it of any change in length of the spindle receptor area.

Since only the primary endings innervate the nuclear bag intrafusal fibers, it is assumed that the nuclear bag fibers are responsible for the powerful dynamic response.

Control of the Static and Dynamic Responses by the Gamma Motor Nerves. Some physiologists believe that the gamma motor nerves to the muscle spindle can be divided into two different types: gamma-dynamic (gamma-d) and gamma-static (gamma-s). The first of these is believed to excite the nuclear bag intrafusal fibers and the second mainly the nuclear chain intrafusal fibers. When the gamma-d fibers excite the nuclear bag fibers, the dynamic response of the muscle spindle becomes tremendously enhanced, whereas the static response is hardly affected. On the other hand, stimulation of the gamma-s fibers, which are believed to excite mainly the nuclear chain fibers, supposedly enhances the static response while having little influence on the dynamic response. We shall see in subsequent paragraphs that these two different types of responses of the muscle spindle are exceedingly important in different types of muscle control.

Continuous Discharge of the Muscle Spindles Under Normal Conditions. Normally, particularly when there is a slight amount of gamma motor excitation, the muscle spindles emit sensory nerve impulses continuously. Stretching the muscle spindles increases the rate of firing, while shortening the spindle decreases this rate of firing. Thus, the spindles can send either *positive signals* to the spinal cord—that is, increased numbers of impulses to indicate increased stretch of a muscle, or they can send *negative signals*—decreased numbers of impulses below the normal level to indicate that the muscle is actually being unstretched.

THE STRETCH REFLEX (ALSO CALLED MYOTATIC REFLEX)

From the prior description of all the intricacies of the muscle spindle, one can readily see that it is a very complex organ. Its function is manifested in the form of the muscle *stretch reflex*—that is, whenever a muscle is stretched, excitation of the spindles causes reflex contraction of the muscle.

Neuronal Circuitry of the Stretch Reflex. Figure 51–4 illustrates the basic circuit of the muscle spindle stretch reflex, showing a type Ia nerve fiber originating in a muscle spindle and entering the dorsal root of the spinal cord. Then, in contrast to most other nerve fibers entering the cord, one branch of it passes directly to the anterior horn of the cord gray matter and synapses directly with anterior motor neurons that send nerve fibers back to the same muscle from whence the muscle spindle fiber originated. Thus, this is a *monosynaptic pathway* that allows a reflex signal to return with almost no delay back to the same muscle after excitation of the spindle.

Some of the type II fibers from the secondary spindle endings also terminate monosynaptically with the anterior motor neurons. However, most of the type II fibers (as well as many collaterals

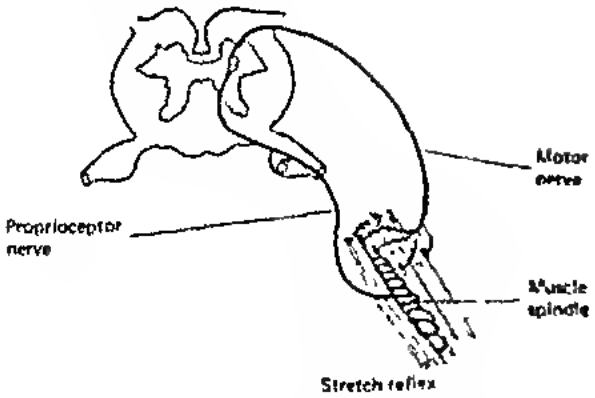


Figure 51–4. Neuronal circuit of the stretch reflex.

from the Ia fibers from the primary endings) terminate on multiple interneurons in the cord gray matter, and these in turn transmit more delayed signals to the anterior motor neurons and also serve other functions.

The Dynamic Stretch Reflex and the Static Stretch Reflex. The stretch reflex can be divided into two separate components called respectively the dynamic stretch reflex and the static stretch reflex. The *dynamic stretch reflex* is elicited by the potent dynamic signal transmitted from the primary endings of the muscle spindles. That is, when a muscle is suddenly stretched, a strong signal is transmitted to the spinal cord, and this causes an instantaneous, very strong reflex contraction of the same muscle from which the signal originated. Thus, the reflex functions to oppose sudden changes in the length of the muscle, because the muscle contraction opposes the stretch.

The dynamic stretch reflex is over within a fraction of a second after the muscle has been stretched to its new length, but then a weaker *static stretch reflex* continues for a prolonged period of time thereafter. This reflex is elicited by the continuous static receptor signals transmitted by both the primary and secondary endings. The importance of the static stretch reflex is that it continues to cause muscle contraction as long as the muscle is maintained at excessive length. The muscle contraction in turn opposes the force that is causing the excess length.

The Negative Stretch Reflex. When a muscle is suddenly shortened, exactly opposite effects occur. If the muscle is already taut, any sudden release of the load on the muscle that allows it to shorten will elicit both dynamic and static reflex muscle inhibition rather than reflex excitation. Thus, *this negative stretch reflex* opposes the shortening of the muscle in the same way that the positive stretch reflex opposes lengthening of the muscle. Therefore, one can begin to see that the stretch reflex tends to maintain the status quo for the length of a muscle.

Feedback Control of Muscle Length—The Load Reflex

Let us assume that a person's biceps is contracted so that the forearm is horizontal to the earth. Then assume that a five-pound weight is put in the hand. The hand will immediately drop. However, the amount that the hand will drop is determined to a great extent by the degree of activity of the muscle spindle reflex. If the gamma-s fibers to the muscle spindles are strongly stimulated so that the static reflex is very active, even slight lengthening of the biceps, and therefore also of the muscle spindles in the biceps, will cause a strong feedback contraction of the extrafusal skeletal muscle fibers of the biceps. This contraction in turn will limit the degree of fall of the hand, thus automatically maintaining the forearm in a nearly horizontal position. This response is called the *load reflex*.

The Damping Function of the Dynamic and Static Stretch Reflexes

Another important function of the stretch reflex—indeed, probably just as important as, if not more important than, the load reflex—is the ability of the stretch reflex to prevent some types of oscillation and jerkiness of the body movements. This is a *damping*, or *smoothing*, function. An example is the following.

Use of the Damping Mechanism in Smoothing Muscle Contraction. Occasionally, signals from other parts of the nervous system are transmitted to a muscle in a very unsmooth form, increasing in intensity for a few milliseconds, then decreasing in intensity, then changing to another intensity level, and so forth. When the muscle spindle apparatus is not functioning satisfactorily, the muscle contraction is very jerky during the course of such a signal. This effect is illustrated in Figure 51–5, which shows an experiment in which a sensory nerve signal entering one side of the cord is transmitted to a motor nerve on the other side of the cord to excite a muscle. In curve A the muscle spindle reflex of the excited muscle is intact. Note that the contraction is relatively smooth even though the sensory nerve is excited at a very slow frequency of 8 per second. Curve B, on the other hand, is the same experiment in an animal whose muscle spindle sensory nerves from the excited muscle had been sectioned three months earlier. Note the very unsmooth muscle contraction. Thus, curve A illustrates very graphically the ability of the damping mechanism of the muscle spindle to smooth muscle contractions even though the input signals to the muscle motor system may themselves be very jerky. This effect can also be called a *signal averaging* function of the muscle spindle reflex.

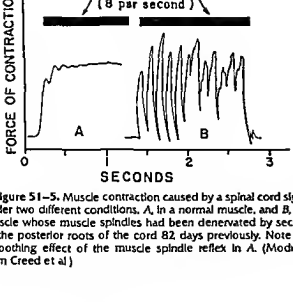


Figure 51–5. Muscle contraction caused by a spinal cord signal under two different conditions. *A*, in a normal muscle, and *B*, in a muscle whose muscle spindles had been denervated by section of the posterior roots of the cord 82 days previously. Note the smoothing effect of the muscle spindle reflex in *A*. (Modified from Creed et al.)

ROLE OF THE MUSCLE SPINDLE IN VOLUNTARY MOTOR ACTIVITY

To emphasize the importance of the gamma efferent system, one needs to recognize that 31 per cent of all the motor nerve fibers to the muscle are gamma efferent fibers rather than large, type A alpha motor fibers. Whenever signals are transmitted from the motor cortex or from any other area of the brain to the alpha motor neurons, almost invariably the gamma motor neurons are stimulated simultaneously, an effect called *co-activation* of the alpha and gamma motor neurons. This causes both the extrafusal and the intrafusal muscle fibers to contract at the same time.

The purpose of contracting the muscle spindle fibers at the same time that the large skeletal muscle fibers contract is probably twofold: First, it keeps the muscle spindle from opposing the muscle contraction. Second, it also maintains proper damping and proper load responsiveness of the muscle spindle regardless of change in muscle length. For instance, if the muscle spindle should not contract and relax along with the large muscle fibers, the receptor portion of the spindle would sometimes be flail and at other times be over-stretched, in neither instance operating under optimal conditions for spindle function.

Possible "Servo" Control of Muscle Contraction by Exciting the Gamma Motor Neurons. Several physiologists have suggested that the muscle spindle reflex can also operate as a servo controller of muscle contraction. A "servo" control system is one in which the output of the system is not activated directly by the input signal but instead is activated indirectly in the following way: Suppose that a gamma efferent nerve fiber to a muscle spindle were stimulated. This would contract the intrafusal fibers and excite the muscle spindle receptors. The signals from these receptors would then pass to the cord and thence back to the extrafusal skeletal muscle fibers to cause them to contract also. Thus, an initial contraction of the spindle intrafusal fibers would cause a similar contraction of the surrounding extrafusal fibers.

If such a servo system as this should exist, signals transmitted from the brain downward through the gamma efferent system could cause secondary contractions of the muscles in accordance with the dictates of the muscle spindles. Unfortunately, most of the data indicate that a mechanism such as this probably does not occur to a significant extent. For a system of this type to function properly, it would be necessary to have a static response of the muscle spindles that is very potent and also unchanging. Unfortunately, the static response of the muscle spindles is usually weak—that is, it has a low gain. Also, it decays very slowly over a period of many minutes to an hour so that changes in length of a muscle spindle could cause *only* temporary changes in length of the surrounding large skeletal muscle fibers. Therefore, despite the attractiveness of this possible type of muscle control, it is doubtful that it does actually occur in normal function.

Probable Servo-Assist Function of the Muscle Spindle Reflex. It is much more likely that the muscle spindle reflex acts as a servo-assist device than as a pure servo controller. But, first, let us explain what is meant by a "servo-assist mechanism."

When both the alpha and gamma motor neurons are stimulated simultaneously, if the intra- and extrafusal fibers contract equal amounts the degree of stimulation of the muscle spindles will *not change at all*—neither increase nor decrease. However, in case the extrafusal muscle fibers should contract less than the intrafusal fibers (as might occur when the muscle is contracting against a great load), this mismatch would stretch the receptor portions of the spindles and, therefore, elicit a stretch reflex that would provide extra excitation of the extrafusal fibers. Thus, failure of the extrafusal fibers to contract to the same degree as the intrafusal fibers contract would cause an accessory neuronal signal that would increase the degree of stimulation of the extrafusal fibers. This is exactly the same mechanism as that employed by power steering in an automobile. That is, if the front wheels are resistant to following the movement of the steering wheel, a servo-assist device becomes activated that applies additional force to turn the wheels.

The servo-assist type of motor function could have several important advantages, as follows.

1. It would allow the brain to cause a muscle contraction against a load without the brain having to expend much extra nervous energy—instead, the spindle reflex would provide most of the nervous energy.

2. It would make the muscle contract almost the *desired length* even when the *load is increased or decreased* between successive contractions. In other words, it would make the length of contraction less loadsensitive.

3. It could compensate for fatigue or other abnormalities of the muscle itself, because any failure of the muscle to provide the proper contraction would elicit an additional muscle spindle reflex stimulus to make the contraction occur.

But, unfortunately, we still do not know how important this probable function of the muscle spindle reflex actually is

Brain Areas for Control of the Gamma Efferent System

The gamma efferent system is excited primarily by the *bulboreticular facilitatory* region of the brain stem, and secondarily by impulses transmitted into this area from (a) the *cerebellum*, (b) the *basal ganglia*, and even (c) the *cerebral cortex*. Unfortunately, though, little is known about the precise mechanisms of control of the gamma efferent system. However, since the bulboreticular facilitatory area is particularly concerned with antigravity contractions, and also because the antigravity muscles have an especially high density of muscle spindles, emphasis is given to the possible or probable important role of the gamma efferent mechanism in controlling muscle contraction for positioning the different parts of the body and for damping the movements of the different parts.

CLINICAL APPLICATIONS OF THE STRETCH REFLEX

The Knee Jerk and Other Muscle Jerks. Clinically, a method used to determine the sensitivity of the stretch reflexes is to elicit the knee jerk and other muscle jerks. The knee jerk can be elicited by simply striking the patellar tendon with a reflex hammer; this stretches the quadriceps muscle and initiates a *dynamic stretch reflex* that in turn causes the lower leg to jerk forward. The upper part of Figure 51–6 illustrates a myogram from the quadriceps muscle recorded during a knee jerk.

Similar reflexes can be obtained from almost any muscle of the body either by striking the tendon of the muscle or by striking the belly of the muscle itself. In other words, sudden stretch of muscle spindles is all that is required to elicit a stretch reflex.

The muscle jerks are used by neurologists to assess the degree of facilitation of spinal cord centers. When large numbers of facilitatory impulses are being transmitted from the upper regions of the central nervous system into the cord, the muscle jerks are greatly exacerbated. On the other hand, if the facilitatory impulses are depressed or abrogated, the muscle jerks are considerably weakened or completely absent. These reflexes are used most frequently to determine the presence or absence of muscle spasticity following lesions in the motor areas of the brain. Ordinarily, large lesions in the contralateral motor areas of the cerebral cortex, especially those caused by strokes or brain tumors, cause greatly exacerbated muscle jerks.

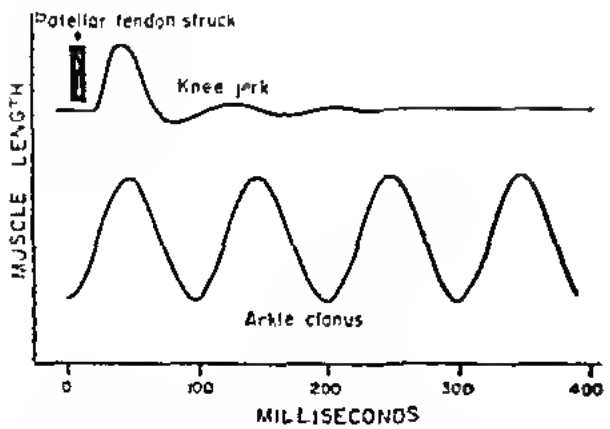


Figure 51–6. Myograms recorded from the quadriceps muscle during elicitation of the knee jerk and from the gastrocnemius muscle during ankle clonus.

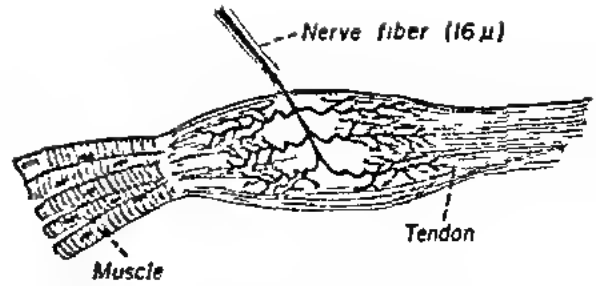


Figure 51–7. Golgi tendon organ.

Clonus. Under appropriate conditions, the muscle jerks can oscillate, a phenomenon called *clonus* (see lower myogram, Fig 51–6). Oscillation can be explained particularly well in relation to ankle clonus, as follows:

If a man standing on his tiptoes suddenly drops his body downward to stretch one of his gastrocnemius muscles, impulses are transmitted from the muscle spindles into the spinal cord. These reflexly excite the stretched muscle, which lifts the body back up again. After a fraction of a second, the reflex contraction of the muscle dies out and the body falls again, thus stretching the spindles a second time. Again a dynamic stretch reflex lifts the body, but this too dies out after a fraction of a second, and the body falls once more to elicit still a new cycle. In this way, the stretch reflex of the gastrocnemius muscle continues to oscillate, often for long periods of time, this is clonus.

Clonus ordinarily occurs only if the stretch reflex is highly sensitized by facilitatory impulses from the brain. For instance, in the decerebrate animal, in which the stretch reflexes are highly facilitated, clonus develops readily. Therefore, to determine the degree of facilitation of the spinal cord, neurologists test patients for clonus by suddenly stretching a muscle and keeping a steady stretching force applied to the muscle. If clonus occurs, the degree of facilitation is certain to be very high.

THE TENDON REFLEX

The Golgi Tendon Organ and Its Excitation. Golgi tendon organs, illustrated in Figure 51–7, are encapsulated sensory receptors through which the muscle tendons pass immediately beyond their attachments to the muscle fibers. An average of 10 to 15 muscle fibers is usually connected in series with each Golgi tendon organ, and the organ is stimulated by the tension produced by this small bundle of muscle fibers. Thus, the major difference between the function of the Golgi tendon organ and the muscle spindle is that the spindle detects relative muscle length, while the tendon organ detects muscle *tension*.

The tendon organ, like the primary receptor of the muscle spindle, has both a *dynamic response* and a *static response*, responding very intensely when the muscle tension suddenly increases (the dynamic response), but within a small fraction of a second it settles down to a lower level of steady

stote firing that is almost directly proportional to the muscle tension (the static response). Thus, the Golgi tendon organs provide the nervous system with instantaneous information of the degree of tension on each small segment of each muscle.

Transmission of Impulses from the Tendon Organ into the Central Nervous System. Signals from the tendon organ are transmitted through large, rapidly conducting type Ib nerve fibers, fibers averaging 16 microns in diameter, only slightly smaller than those from the primary ending of the muscle spindle. These fibers, like those from the primary endings, transmit signals both into local areas of the cord and through the spinocerebellar tracts into the cerebellum and through still other tracts to the cerebral cortex. The local cord signal excites a single inhibitory interneuron that in turn inhibits the anterior alpha motor neuron. This local circuit directly inhibits the individual muscle without affecting adjacent muscles. The signals to the brain will be discussed in Chapter 53.

Inhibitory Nature of the Tendon Reflex. When the Golgi tendon organs of a muscle are stimulated by increased muscle tension, signals are transmitted into the spinal cord to cause reflex effects in the respective muscle. However, this reflex is entirely inhibitory, the exact opposite to the muscle spindle reflex. Thus, this reflex provides e negative feedback mechanism that prevents the development of too much tension on the muscle.

When tension on the muscle and, therefore, on the tendon becomes extreme, the inhibitory effect from the tendon organ can be so great that it causes sudden relaxation of the entire muscle. (The signals from the secondary endings of the muscle spindles are believed also to help in causing this reaction.) This effect is called the *lengthening reaction;* it is probably a protective mechanism to prevent tearing of the muscle or avulsion of the tendon from its attachments to the bone. We know, for instance, that direct electrical stimulation of muscles in the laboratory, which cannot be opposed by this tension reflex, can frequently cause such destructive effects.

The Tendon Reflex as a Possible Servo Control Mechanism for Muscle Tension. In the same way that the stretch reflex possibly operates as a feedback mechanism to control muscle length, the tendon reflex theoretically can operate as a servo feedback mechanism to control muscle tension. That is, if the tension on the muscle becomes too great, inhibition from the tendon organ decreases this tension back to a lower value. On the other hand, if the tension becomes too little, impulses from the tendon organ cease; and the resulting loss of inhibition allows the alpha motor neurons to become active again, thus increasing muscle tension back toward a higher level.

Very little is known at present about the function or control of this tension servo feedback mechanism, but it is postulated to operate in the following basic manner: Signals from the brain are presumably transmitted to the cord to apprise a neuronal group in the cord of the tension required in each given muscle. At the same time, signals are also transmitted through the alpha motor neurons to cause muscle contraction. Then, as the degree of contraction approaches the tension limit set for the muscle, as detected by the feedback from the Golgi tendon organs, the neuronal group automatically inhibits the muscle contraction to prevent additional tension. In this way, the tension becomes adjusted to the "set-point" dictated by the brain.

Yet, a word of caution: Despite this speculation, the measured feedback gains of this muscle tension feedback system have been far too little to prove that it can be very useful as a tension servo controller.

Possible Value of a Servo Mechanism for Control of Tension. An obvious value of a mechanism for setting the degree of muscle tension would be to allow the different muscles to apply a desired amount of force irrespective of how far the muscles contract. An example of this is paddling a boat, in which a person sets the amount of force with which he pulls backward on the paddle and maintains that degree of force throughout the entire movement. Were it not for some type of tension feedback mechanism, the same amount of force would not be maintained throughout the stroke because uncontrolled muscles change their force of contraction as their lengths change. One can imagine hundreds of different patterns of muscle function that might require maintenance of constant tension rather than maintenance of constant lengths of the muscles.

THE FLEXOR REFLEX (THE WITHDRAWAL REFLEXES)

In the spinal or decerebrate animal, almost any type of cutaneous sensory stimulus on a limb is likely to cause the flexor muscles of the limb to contract, thereby withdrawing the limb from the stimulus. This is called the flexor reflex.

In its classic form the flexor reflex is elicited most powerfully by stimulation of pain endings, such as by pinprick, heat, or some other painful stimulus, for which reason it is also frequently called a *nociceptive reflex,* or simply *pain reflex.* However, even stimulation of the touch receptors can also occasionally elicit a weaker and less prolonged flexor reflex.

If some part of the body besides one of the limbs is painfully stimulated, this part, in a similar manner, will be withdrawn from the stimulus, but the reflex may not be confined entirely to flexor muscles even though it is basically the same type of reflex. Therefore, the many patterns of reflexes of this type in the different areas of the body are called the *withdrawal reflexes.*

Neuronal Mechanism of the Flexor Reflex. The left-hand portion of Figure 51–8 illustrates the neuronal pathways for the flexor reflex. In this instance, a painful stimulus is applied to the hand; as a result, the flexor muscles of the upper arm become reflexly excited, thus withdrawing the hand from the painful stimulus.

The pathways for eliciting the flexor reflex do not pass directly to the anterior motor neurons but, instead, pass first into the interneuron pool of neurons and then to the motor neurons. The shortest possible circuit is a three- or four-neuron arc; however, most of the signals of the reflex traverse

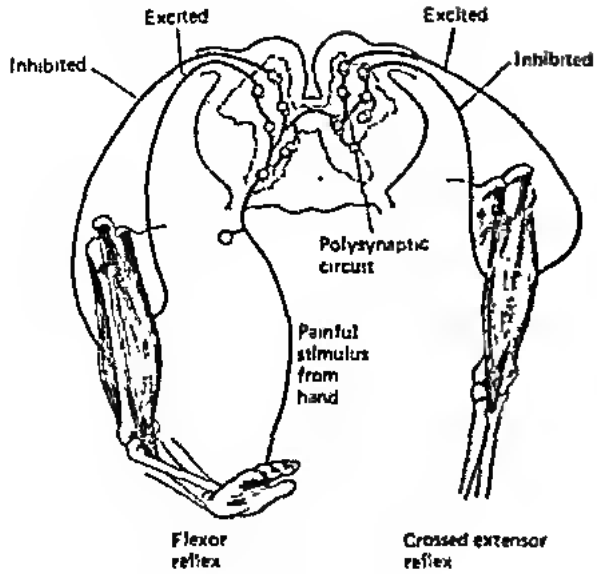


Figure 51–8. The flexor reflex, the crossed extensor reflex, and reciprocal inhibition.

many more neurons than this and involve the following basic types of circuits: (1) diverging circuits to spread the reflex to the necessary muscles for withdrawal, (2) circuits to inhibit the antagonist muscles, called *reciprocal inhibition circuits*, and (3) circuits to cause a prolonged repetitive after-discharge even after the stimulus is over.

Figure 51–9 illustrates a typical myogram from a flexor muscle during a flexor reflex. Within a few milliseconds after a pain nerve begins to be stimulated, the flexor response appears. Then, in the next few seconds the reflex begins to *fatigue*, which is characteristic of essentially all of the more complex integrative reflexes of the spinal cord. Then, soon after the stimulus is over, the contraction of the muscle begins to return toward the base line, but, because of *after-discharge*, will not return all the way for many milliseconds. The duration of the after-discharge depends on the intensity of the sensory stimulus that had elicited the reflex; a weak tactile stimulus causes almost no after-discharge in contrast to an after-discharge lasting for several seconds following a very strong pain stimulus.

The after-discharge that occurs in the flexor

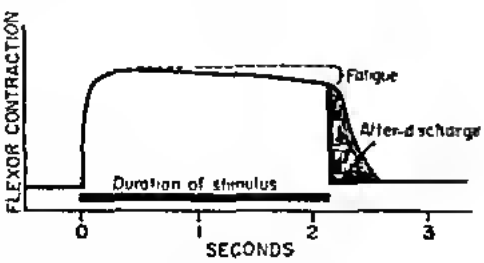


Figure 51–9. Myogram of the flexor reflex, showing rapid onset of the reflex, an interval of fatigue, and finally after-discharge after the stimulus is over.

reflex almost certainly results from all three types of repetitive-discharge circuits that were discussed in Chapter 47. Electrophysiological studies indicate that the immediate after-discharge, lasting for about 6 to 8 milliseconds, results from the interneuron repetitive firing mechanism and from the parallel type of circuit, with impulses being transmitted from one interneuron to another to another and all these in turn transmitting their signals successively to the anterior motor neurons. However, the prolonged after-discharge that occurs following strong pain stimuli presumably involves reverberating circuits in the interneurons, these transmitting impulses to the anterior motor neurons sometimes for several seconds after the incoming sensory signal is completely over.

Thus, the flexor reflex is appropriately organized to withdraw a pained or otherwise irritated part of the body away from the stimulus. Furthermore, because of the after-discharge the reflex can still hold the irritated part away from the stimulus for as long as 1 to 3 seconds after the irritation is over. During this time, other reflexes and actions of the central nervous system can move the entire body away from the painful stimulus.

The Pattern of Withdrawal. The pattern of withdrawal that results when the flexor reflex (or the many other types of withdrawal reflexes) is elicited depends on the sensory nerve that is stimulated. Thus, a painful stimulus on the inside of the arm not only elicits a flexor reflex in the arm but also contracts the abductor muscles to pull the arm outward. In other words, the integrative centers of the cord cause those muscles to contract that can most effectively remove the pained part of the body from the object that causes pain. This same principle, which is called the principle of "local sign," applies for any part of the body but especially to the limbs, because they have highly developed flexor reflexes.

THE CROSSED EXTENSOR REFLEX

Approximately 0.2 to 0.5 second after a stimulus elicits a flexor reflex in one limb, the opposite limb begins to extend. This is called the *crossed extensor reflex*. Extension of the opposite limb obviously can push the entire body away from the object causing the painful stimulus.

Neuronal Mechanism of the Crossed Extensor Reflex. The right-hand portion of Figure 51–8 illustrates the neuronal circuit responsible for the crossed extensor reflex, showing that signals from the sensory nerves cross to the opposite side of the cord to cause exactly opposite reactions to those that cause the flexor reflex. Because the crossed extensor reflex usually does not begin until 200 to 500 milliseconds following the initial pain stimulus, it is certain that many interneurons are in the circuit between the incoming sensory neuron and the motor neurons of the opposite side of the

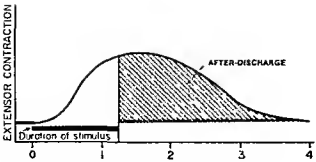


Figure 51–10. Myogram of a crossed extensor reflex, showing slow onset but prolonged after-discharge.

cord responsible for the crossed extension. Furthermore, after the painful stimulus is removed, the crossed extensor reflex continues for an even longer period of after-discharge than that for the flexor reflex. Therefore, again, it is presumed that this prolonged after-discharge results from reverberatory circuits among tha internuncial cells

Figure 51–10 illustrates a myogram recorded from a muscle involved in a crossed extensor reflex. This shows the relatively long latency before the reflex begins and also tho long after-discharge following the end of the stimulus. The prolonged after-discharga obviously would be of benefit in bolding the body away from a painful object until other nervous reactions could cause the body to move away.

RECIPROCAL INHIBITION AND RECIPROCAL INNERVATION

In the foregoing paragraphs we have pointed out several times that excitation of one group of muscles is often associated with inhibition of another group. For instance, when a stretch reflex excites one muscle, it simultaneously inhibits the antagonist muscles. This is the phenomenon of *reciprocal inhibition,* and the neuronal mechanism that causes this reciprocal relationship is called *reciprocal innervation.* Likewise, reciprocal relationships exist between the two sides of the cord, as exemplified by the flexor and extensor reflexes described above.

Figure 51–11 illustrates a typical example of reciprocal inhibition. In this instance, a moderate

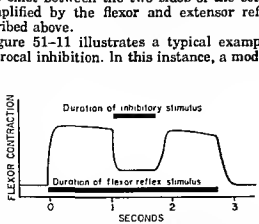


Figure 51–11. Myogram of a flexor reflex, illustrating reciprocal inhibition caused by a stronger flexor reflex in the opposite limb.

but prolonged flexor reflex is elicited from one limb of the body, and while this reflex is still being elicited a still stronger flexor reflex is elicited in the opposite limb, causing reciprocal inhibition of the first limb. Then removal of the strong flexor reflex allows the original reflex to reassume its previous intensity.

FATIGUE OF REFLEXES; REBOUND

Figure 51–9 illustrated that the flexor reflex begins to *fatigue* within a few seconds after its initiation. This is a common effect in most of the cord reflexes as well as many other reflexes of the central nervous aystem, and it presumably results from progressive fatigue of synaptic transmission in the reflex circuits, a subject that was discussed more fully in Cbapter 46.

Another effect closely allied to fatigue is *rebound.* This means that, immediately after a reflex ia over, a second reflex of the same type is much more difficult to elicit for a given time thereafter. However, because of reciprocal innervation, reflexes of the antagonist muscles become even more easily elicited during that same period of time. For instance, if a flexor reflex occurs in a left limb, a second flexor reflex is more difficult to establish for a few seconds thereafter, but a crossed extensor reflex in this same limb will ba greatly exacerbated. Rebound is probably one of the important mechanisms for coordinating tha rhythmic to-and-fro movements required in locomotion, described in more detail later in the chapter.

THE REFLEXES OF POSTURE AND LOCOMOTION

THE POSTURAL AND LOCOMOTIVE REFLEXES OF THE CORD

The Positive Supportive Reaction. Pressure on the footpad of a decerebrate animal causes the limb to extend against the pressure being applied to the foot. Indeed, this reflex is so strong that an animal whose spinal cord has been transected for several months, when the reflexes have become exacerbated, can often be placed on its feet, and the reflex will stiffen the limbs sufficiently to support the weight of the body—the animal will stand in a rigid position. This reflex is called the *positive supportive reaction.*

The positive supportive reaction involves a complex circuit in the interneurons similar to those responsible for the flexor and the crossed-extensor reflexes. Furthermore, the locus of the pressure on the pad of the foot determines the position to which the limb will extend; pressure on one side causes extension in that direction, an effect called the *magnet reaction.* This obviously helps keep an animal from falling to that side.

The Cord "Righting" Reflexes. When a spinal cat or even a well-recovered young spinal dog is laid on its side, it will make incoordinate movements that indicate that it is trying to raise itself to the standing position. This is called a *cord righting reflex*, and it illustrates that relatively complicated reflexes associated with posture are at least partially integrated in the spinal cord. Indeed, a puppy with a well-healed transected thoracic cord between the cord level for the forelimbs and the hindlimbs can completely right itself from the lying position and can even walk on its hindlimbs. And, in the case of the opossum with a similar transection of the thoracic cord, the walking movements of the hindlimbs are hardly different from those in the normal opossum—except that the hindlimb movements are not synchronized with those of the forelimbs as is normally the case.

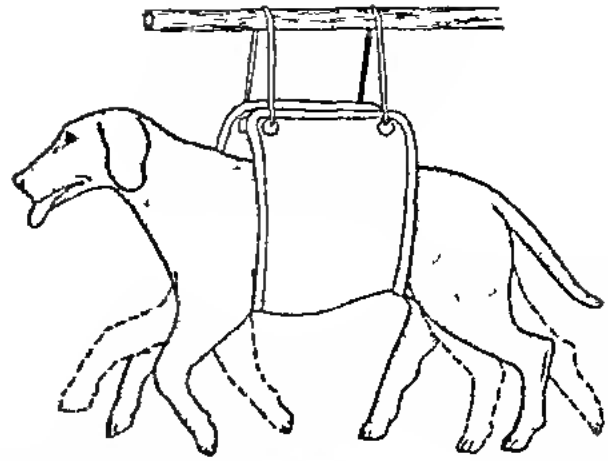


Figure 51–12. Diagonal stepping movements exhibited by a spinal animal

Stepping and Walking Movements

Rhythmic Stepping Movements of a Single Limb. Rhythmic stepping movements are frequently observed in the limbs of spinal animals. Indeed, even when the lumbar portion of the spinal cord is separated from the remainder of the cord and a longitudinal section is made down the center of the cord to block neuronal connections between the two limbs, each hind limb can still perform stepping functions. Forward flexion of the limb is followed a second or so later by backward extension. Then flexion occurs again, and the cycle is repeated over and over.

This oscillation back and forth between the flexor and extensor muscles occurs even after the sensory nerves have been cut, and it seems to result mainly from reciprocal inhibition and rebound circuits that oscillate within the matrix of the cord itself. That is, the forward flexion of the limb causes reciprocal inhibition of the cord center controlling the extensor muscles, but shortly thereafter the flexion begins to die out, as it does so, *rebound* inhibition of the flexors and reciprocal *rebound* excitation of the extensors cause the leg to move downward and backward. After extension has continued for a time, it, too, begins to die and is followed by reciprocal rebound excitation of the flexor muscles. And the oscillating cycle continues again and again.

The sensory signals from the footpads and from the position sensors around the joints play a strong role in controlling foot pressure and rate of stepping when the foot is allowed to walk along a surface. In fact, the cord mechanism for control of stepping can be still more complex. For instance, if during the forward thrust of the foot the top of the foot encounters an obstruction, the forward thrust will stop temporarily, the foot will be lifted higher, and then the foot will proceed forward to be placed over the obstruction. Thus, the cord is an intelligent walking controller.

Reciprocal Stepping of Opposite Limbs. If the lumbar spinal cord is not sectioned down its center as noted above, every time stepping occurs in the forward direction in one limb, the opposite limb ordinarily steps backward. This effect results from reciprocal innervation between the two limbs.

Diagonal Stepping of All Four Limbs—The "Mark Time" Reflex. If a well-healed spinal animal, with the spinal transection above the forelimb area of the cord, is held up from the table and its legs are allowed to fall downward as illustrated in Figure 51–12, the stretch on the limbs occasionally elicits stepping reflexes that involve all four limbs. In general, stepping occurs diagonally between the fore- and hindlimbs. That is, the right hindlimb and the left forelimb move backward together while the right forelimb and left hindlimb move forward. This diagonal response is another manifestation of reciprocal innervation, this time occurring the entire distance up and down the cord between the fore- and hindlimbs. Such a walking pattern is often called a *mark time reflex*.

The Galloping Reflex. Another type of reflex that occasionally develops in the spinal animal is the galloping reflex, in which both forelimbs move backward in unison while both hindlimbs move forward. If stretch or pressure stimuli are applied almost exactly equally to opposite limbs at the same time, a galloping reflex will likely result, whereas unequal stimulation of one side versus the other elicits the diagonal walking reflex. This is in keeping with the normal patterns of walking and of galloping, for, in walking, only one limb at a time is stimulated, and this would predispose to continued walking. Conversely, when the animal strikes the ground during galloping, the limbs on both sides are stimulated approximately equally; this obviously would predispose to further galloping and, therefore, would continue this pattern of motion in contradistinction to the walking pattern.

THE SCRATCH REFLEX

An especially important cord reflex in lower animals is the scratch reflex, which is initiated by the *itch and tickle sensation*. It actually involves two different functions. (1) a *position sense* that allows the paw to find the exact point of irritation on the surface of the body, and (2) a *to-and-fro scratching movement*

Obviously, the scratch reflex, like the stepping movements of locomotion, involves reciprocal innervation circuits that cause oscillation. One of the important discoveries in relation to the to-and-fro movement of the scratch reflex (and also of the stepping movements as described above) is that it can still occur even when all the sensory roots from the oscillating limb are sectioned. In other words, feedback from the limb itself is not necessary to maintain the neuronal oscillation, which means that the oscillation can occur intrinsically as a result of oscillating circuits within the spinal interneurons themselves

The *position sense* of the scratch reflex is also a highly developed function, for even though a flea might be crawling as far forward as the shoulder of a spinal animal, the hind paw can often find its position Furthermore, this can be accomplished even though 19 different muscles in the limb must be contracted simultaneously in a precise pattern to bring the paw to the position of the crawling flea To make the reflex even more complicated, when the flea crosses the midline, the paw stops scratching, but the opposite paw begins the to-and-fro motion and eventually finds the flea

THE SPINAL CORD REFLEXES THAT CAUSE MUSCLE SPASM

In human beings, local muscle spasm is often observed. The mechanism of this has not been elucidated to complete satisfaction even in experimental animals, but it is known that pain stimuli can cause reflex spasm of local muscles, which presumably is the cause of much if not most of the muscle spasm *observed at localized* regions of the human body.

Muscle Spasm Resulting from a Broken Bone. One type of clinically important spasm occurs in muscles surrounding a broken bone. This seems to result from the pain impulses initiated from the broken edges of the bone, which cause the muscles surrounding the area to contract powerfully and tonically. Relief of the pain by injection of a local anesthetic relieves the spasm; a general anesthetic also relieves the spasm. One of these procedures is often necessary before the spasm can be overcome sufficiently for the two ends of the bone to be set back into appropriate positions.

Abdominal Muscle Spasm in Peritonitis. Another type of local spasm caused by a cord reflex is the abdominal spasm resulting from irritation of the parietal peritoneum by peritonitis. Here, again, relief of the pain caused by the peritonitis allows the spastic muscle to relax. Almost the same type of spasm often occurs during surgical operations; pain impulses from the parietal peritoneum cause the abdominal muscles to contract extensively and sometimes actually to extrude the intestines through the surgical wound. For this reason deep surgical anesthesia is usually required for intra-abdominal operations.

Muscle Cramps. Still another type of local spasm is the typical muscle cramp Electromyographic studies indicate that the cause of at least some muscle cramps is the following:

Any local irritating factor or metabolic abnormality of a muscle—such as severe cold, lack of blood flow to the muscle, or overexercise of the muscle—can elicit pain or other types of sensory impulses that are transmitted from the muscle to the spinal cord, thus causing reflex muscle contraction. The contraction in turn stimulates the same sensory receptors still more, which causes the spinal cord to increase the intensity of contraction still further. Thus, a positive feedback mechanism occurs so that a small amount of initial irritation causes more and more contraction until a full-blown muscle cramp ensues Reciprocal inhibition of the muscle can sometimes relieve the cramp. That is, if a person purposefully contracts the muscle on the opposite side of the joint from the cramped muscle while at the same time using another hand or foot to prevent movement of the joint, the reciprocal inhibition that occurs in the cramped muscle can at times relieve the cramp

THE AUTONOMIC REFLEXES IN THE SPINAL CORD

Many different types of segmental autonomic reflexes occur in the spinal cord, most of which are discussed in other chapters Briefly, these include (1) changes in vascular tone, resulting from local skin heat and cold (Chapter 72), (2) sweating, which results from localized heat on the surface of the body (Chapter 72), (3) intestinointestinal reflexes that control some motor functions of the gut (Chapter 63), (4) peritoneointestinal reflexes that inhibit gastric motility in response to peritoneal irritation (Chapter 66), and (5) evacuation reflexes for emptying the bladder (Chapter 38) and the colon (Chapter 63). In addition, all the segmental reflexes can at times be elicited simultaneously in the form of the so-called mass reflex as follows

The Mass Reflex. In a spinal animal or human being, the spinal cord sometimes suddenly becomes excessively active, causing massive discharge of large portions of the cord The usual stimulus that causes this is a strong nociceptive stimulus to the skin or excessive filling of a viscus, such as overdistention of the bladder or of the gut. Regardless of the type of stimulus, the resulting reflex, called the *mass reflex*, involves large portions or even all of the cord, and its pattern of reaction is the same. The effects are (1) a major portion of the body goes into strong flexor spasm, (2) the colon and bladder are likely to evacuate, (3) the arterial pressure often rises to maximal values—sometimes to a mean pressure well over 200 mm Hg, and (4) large areas of the body break out into profuse sweating. The mass reflex might be likened to the epileptic seizures that involve the central nervous system in which large portions of the brain become massively activated.

The precise neuronal mechanism of the mass reflex is unknown. However, since it lasts for minutes, it presumably results from activation of great masses of reverberating circuits that excite large areas of the cord at once.

SPINAL CORD TRANSECTION AND SPINAL SHOCK

When the spinal cord is suddenly transected, essentially all cord functions, including [illegible] reflex

immediately become depressed to the point of oblivion, a reaction called *spinal shock*. The reason for this is that normal activity of the cord neurons depends to a great extent on continual tonic discharges from higher centers, particularly discharges transmitted through the reticulospinal tracts, vestibulospinal tracts, and corticospinal tracts.

After a few hours to a few days or weeks of spinal shock, the spinal neurons gradually regain their excitability. This seems to be a natural characteristic of neurons everywhere in the nervous system—that is, after they lose their source of facilitatory impulses, they increase their own natural degree of excitability to make up for the loss. But there is also some possibility of sprouting of multiple new nerve endings in the cord, which could also increase excitability. In most nonprimates, the excitability of the cord centers returns essentially to normal within a few hours to a day or so, but in human beings the return is often delayed for several weeks and occasionally is never complete; or, on the other hand, recovery is sometimes excessive, with resultant hyperexcitability of all or most cord functions

Some of the spinal functions specifically affected during or following spinal shock are (1) The arterial blood pressure falls immediately—sometimes to as low as 40 mm Hg—thus illustrating that sympathetic activity becomes blocked almost to extinction. However, the pressure ordinarily returns to normal within a few days (2) All skeletal muscle reflexes integrated in the spinal cord are completely blocked during the initial stages of shock In lower animals, a few hours to a few days are required for these reflexes to return to normal, and in human beings two weeks to several months are usually required. Sometimes, both in animals and people, some reflexes eventually become hyperexcitable, particularly if a few facilitatory pathways remain intact between the brain and the cord while the remainder of the spinal cord is transected. The first reflexes to return are the stretch reflex, followed in order by the progressively more complex reflexes, the flexor reflexes, the postural antigravity reflexes, and remnants of stepping reflexes (3) The sacral reflexes for control of bladder and colon evacuation are completely suppressed in human beings for the first few weeks following cord transection, but *they eventually return These effects are discussed in* Chapters 38 and 66.

REFERENCES

Arbib, M A Perceptual structures and distributed motor control *In* Brooks, V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 1449

Austin, G The Spinal Cord New York, Igaku Shoin Medical Publishers, 1981

Baldissera, F, Hultborn, H, and Illert, M · Integration in spinal neuronal systems *In* Brooks, V B (ed.) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 509

Burke, R. E., Motor units Anatomy, physiology, and functional organization *In* Brooks, V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 345

Burke, R. E, and Rudomin, P., Spinal neurons and synapses *In* Brookhart, J M, and Mountcastle, V B (eds). Handbook of Physiology Sec. 1, Vol 1 Baltimore Williams & Wilkins, 1977 p 877

Creed, R S, *et al* Reflex Activity of the Spinal Cord New York, Oxford University Press, 1932

Desmedt, J E (ed) Physiological Tremor, Pathological Tremors and Clonus. New York, S Karger, 1978

Easton, T. A., On the normal use of reflexes. *Am Sci*, *60* 591, 1972

Freund, H J., Motor unit and muscle activity in voluntary motor control *Physiol Rev*, *63* 387, 1983

Gallistel, C. R The Organization of Action A New Synthesis. New York, Halsted Press, 1979

Granit, R · Muscular tone, *J. Sport Med*, *2* 46, 1962

Granit, R., Comments on history of motor control. *In* Brooks, V. B (ed) Handbook of Physiology. Sec 1, Vol II. Bethesda, American Physiological Society, 1981, p 1

Granit R, and Pompeiano, O (eds.) Reflex Control of Posture and Movement. New York, Elsevier Scientific Publishing Co, 1979

Granit, R., *et al*: First supraspinal control of mammalian muscle spindles, extra- and intrafusal co-activation *J Physiol (Lond)* *147*:385, 1959

Grillner, S Control of locomotion in bipeds, tetrapods, and fish *In* Brooks, V B (ed) Handbook of Physiology Sec. 1, Vol II Bethesda, American Physiological Society, 1981, p 1179

Grillner, S. Neurobiological bases of rhythmic motor acts in vertebrates. Science *228* 143, 1985

Halstead, L. S. Neuractive Drugs of Choice in Spinal Cord Injury A Guide for Using Neurologically Active Medications in Spinal Injured Patients New York, Raven Press, 1981

Henneman, E, and Mendell, L M Functional organization of motoneuron pool and its inputs *In* Brooks, V B (ed.) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981 p 423

Houk, J C Regulation of stiffness in skeletomotor reflexes *Annu Rev Physiol*, *41* 99, 1979

Houk, J C, and Rymer, W Z. Neural control of muscle length and tension *In* Brooks, V B (ed) Handbook of Physiology Sec. 1, Vol II Bethesda, American Physiological Society, 1981, p 257

Hughes, J T Pathology of the Spinal Cord Philadelphia, W. B Saunders Co., 1978

Hunt, C C, and Perl, E R Spinal reflex mechanisms concerned with skeletal muscle *Physiol Rev*, *40*:538, 1960

Kao C C (ed) Spinal Cord Reconstruction. New York, Raven Press, 1983

Kostyuk, P G, and Vasilenko, D Spinal interneurons. *Annu Rev Physiol* *41* 115, 1979

Le Douarin, N M, *et al* From the neural crest to the ganglia of the peripheral nervous system *Annu Rev Physiol*, *43*:653, 1981

Matthews, P B C Mammalian Muscle Receptors and Their Central Actions Baltimore, Williams & Wilkins, 1972

Matthews, P B C Muscle spindles. Their messages and their fusimotor supply *In* Brooks V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 189

Mendell, L. M Modifiability of spinal synapses *Physiol Rev*, *64* 260, 1984

Merton, P A How we control the contraction of our muscles. *Sci. Am.*, *226*:30 1972

Orlovsky, G N, and Shik, M L. Control of locomotion. A neurophysiological analysis of the cat locomotor system. *Int Rev Physiol*, *10* 281, 1976

Pearson, K The control of walking *Sci Am* *235*(6) 72, 1976

Peterson, B W Reticulospinal projections to spinal motor nuclei *Annu Rev Physiol* *41* 127, 1979

Porter, R. The neurophysiology of movement performance *In MTP* International Review of Science Physiology Vol 3 Baltimore, University Park Press, 1974 p 151

Purves, D Long term regulation in the vertebrate peripheral nervous system *Int Rev Physiol*, *10*:125, 1976

Rack, P M. H Limitations of somatosensory feedback in control of posture and movement *In* Brooks, V B (ed). Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 229

Rogers, M. A. Paraplegia, Boston, Faber and Faber, 1978.

Rowell, L. B, Reflex control of regional circulation in humans *J. Auton Nerv Syst*, *11* 101, 1984

Sherrington, C S The Integrative Action of the Nervous System. New Haven, Conn, Yale University Press, 1911

Shik, M L., and Orlovsky, G N: Neurophysiology of Locomotor Automatism *Physiol Rev*, *56* 465, 1976

Stein, P S G., Motor systems with specific reference to the control of locomotion *Annu Rev. Neurosci*, *1* 61, 1978

Stein, R. B Peripheral control of movement. *Physiol Rev*, *54* 215, 1974

Stein, R. B, and Lee, R. G · Tremor and clonus. *In* Brooks, V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 325

Wiesendanger, M, and Miles, T S · Ascending pathway of low-threshold muscle afferents to the cerebral cortex and its possible role in motor control *Physiol Rev*, *62* 1234, 1982

Young, R: Spinal Cord Injuries Case Studies New Hyde Park, N Y, Medical Examination Publishing Co, 1981

52

Motor Functions of the Brain Stem and Basal Ganglia—Reticular Formation, Vestibular Apparatus, Equilibrium, and Brain Stem Reflexes

The brain stem is a complex extension of the spinal cord. Collected in it are numerous neuronal circuits to control respiration, cardiovascular function, gastrointestinal function, eye movement, equilibrium, support of the body against gravity, and many special stereotyped movements of the body. Some of these functions—such as control of respiration and cardiovascular functions—are described in special sections of this text. In the present chapter we shall discuss the role of the brain stem in controlling whole body movement and equilibrium.

THE RETICULAR FORMATION, AND SUPPORT OF THE BODY AGAINST GRAVITY

Throughout the entire extent of the brain stem—in the medulla, pons, and mesencephalon—are areas of diffuse neurons collectively known as the *reticular formation*. Figure 52–1 illustrates the extent of the reticular formation, showing it to begin at the upper end of the spinal cord and to extend upward through the medulla, pons, and mesencephalon. Many of the ascending and descending fiber tracts between the brain and spinal cord pass through the reticular formation, and as they do so they also provide collateral nerve endings to all reticular areas. In turn, the reticular formation provides multiple efferent fibers that pass both upward and downward in the axis of the nervous system. The lower end of the reticular formation is continuous with the interneurons of the spinal cord, and, indeed, the reticular formation of the brain stem functions in a manner quite

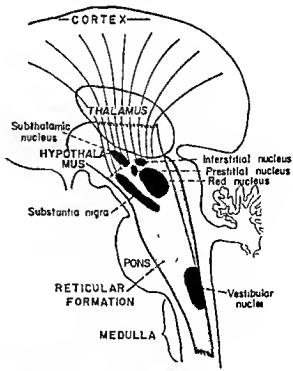


Figure 52–1. The reticular formation and

analogous to many of the functions of the interneurons in the cord gray matter.

Interspersed throughout the reticular formation are both motor and sensory neurons; these vary in size from very small to very large. The small neurons, which constitute the greater number, make multiple connections within the reticular formation itself. The large neurons are mainly motor in function, and their axons often bifurcate almost immediately, with one division extending downward to the spinal cord and the other extending upward to the thalamus or other basal regions of the diencephalon or cerebrum. The axons passing into the spinal cord help control the axial and girdle movements—that is, the trunk, neck, shoulder, hip, and proximal limb movements. On the other hand, the nerve fibers extending into the diencephalon and cerebrum play important roles in controlling overall activity of the brain or activity in specific portions of the brain. These especially play prominent roles in the control of wakefulness and sleep, as we shall discuss in Chapter 55.

The input signals to the reticular formation are derived from multiple sources, including: (1) the spinoreticular tracts and collaterals from the spinothalamic tracts, (2) the vestibular nuclei, (3) the cerebellum, (4) the basal ganglia, (5) the cerebral cortex, from both the sensory and the motor regions, and (6) the hypothalamus and other nearby associated areas.

Though most of the neurons in the reticular formation are evenly dispersed, some of them are collected into *specific nuclei*, some of which are labeled in Figure 52–1. In general, these specific nuclei are not considered to be part of the reticular formation per se even though they do operate in association with it. In most instances they are the loci of "preprogrammed" control of stereotyped movements, as we shall discuss later in the chapter. Note especially in Figure 52–1 the location of the vestibular nuclei in the posterior portion of the lower pons and upper medulla. These nuclei provide preprogrammed attitudinal contractions of appropriate muscles for maintaining equilibrium, and they share almost equally with the reticular formation in the control of the other motor functions by the brain stem.

The Motor Pathways from the Brain Stem to the Spinal Cord. Figure 52–2 illustrates the principal motor pathways from the brain stem to the spinal cord. These include

(1) Two vestibulospinal tracts that originate in the vestibular nuclei: a *lateral vestibulospinal tract* and a *medial vestibulospinal tract*. Both these are located in the anteromedial column of the spinal cord, and they excite motor neurons located mainly in the medial portion of the anterior horn. These neurons control the axial and girdle muscles.

(2) Two reticulospinal tracts: a *pontine reticulospinal tract* located in the anteromedial column of the cord, and a *medullary reticulospinal tract* located in the lateral column. The pontine tract originates principally in the

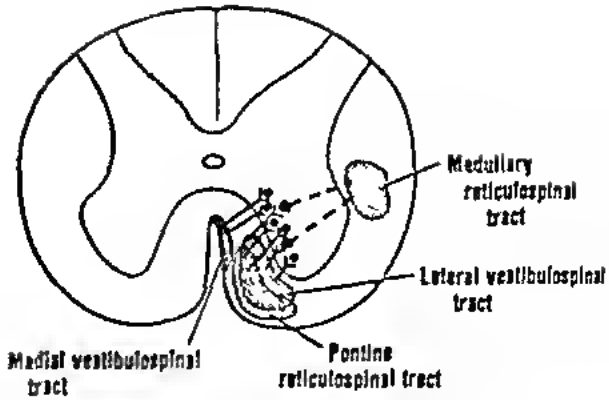


Figure 52–2. The vestibulospinal and reticulospinal tracts descending in the spinal cord to excite (solid lines) or inhibit (dashed lines) the anterior motor neurons that control the body's axial musculature.

reticular formation of the pons whereas the medullary tract originates in the medulla. Both terminate on the medial motor neurons of the anterior horn. However, the signals from the pontine reticulospinal tract are mainly excitatory whereas those from the medullary tract are mainly inhibitory.

(3) Imbedded within the lateral vestibulospinal tract and the pontine reticulospinal tract are several smaller tracts from the brain stem, including the *tectospinal tract*, which originates in the tectum of the mesencephalon, and the *interstitiospinal* tract, which originates in the interstitial nucleus within the reticular formation of the mesencephalon.

Note that most of the pathways from the brain stem to the spinal cord are located in the anteromedial columns of the cord and that they terminate on the medial motor neurons of the anterior horn. These motor neurons control the axial musculature of the body and the musculature of the two body girdles—the shoulders and hips and proximal portions of the limbs. This complex of pathways and motor neurons is sometimes called the *medial motor system;* it is important because it is the principal controller of background body and girdle muscle contractions. It especially provides contraction of the antigravity muscles to support the body against gravity, while inhibiting the flexor muscles to prevent their opposition to the antigravity muscles. However, the medullary reticulospinal tract is different from the other tracts originating in the brain stem, for it tends to inhibit the antigravity muscles while providing perhaps some degree of excitation to the flexors.

Excitation of the Antigravity Muscles by the Brain Stem—The Decerebrate Animal

Both the vestibular nuclei and the pontine portion of the reticular formation, the two of which function in very close association with each other, are intrinsically excitable. However, this excitability is usually held in check by inhibitory signals that originate mainly in the basal ganglia but to a lesser extent in the cerebral cortex and cerebellum. When the brain stem is sectioned above the vestibular nuclei at the midpontine level, the intrinsic activity of these nuclei immediately causes a considerable degree of rigidity of all the anti-

gravity muscles in the body. If this section is made still higher in the brain stem, for instance at the midlevel of the mesencephalon, then the intrinsic activity of the pontine portion of the reticular formation adds still more to the rigidity of the antigravity muscles, and the animal can usually even stand on its feet. This is called the *decerebrate animal.* Finally, if the cerebellum is removed at the same time to remove its inhibitory influences, the rigidity becomes still further enchanced.

Thus, the principal role of the brain stem in the control of motor function is to provide background contractions of the trunk and neck musculature and proximal portions of the limbs; and the most important function of these contractions is to provide support of the body against gravity.

The relative degree of contraction of the individual antigravity muscles is determined by the equilibrium mechanisms. Thus, if an animal begins to fall to one side, the extensor muscles on that side stiffen while those on the opposite side relax. In turn, the equilibrium mechanisms are controlled by signals mainly from the *vestibular apparatuses,* which are the sensory organs for equilibrium that will be discussed fully in subsequent sections of this chapter.

VESTIBULAR SENSATIONS AND THE MAINTENANCE OF EQUILIBRIUM

THE VESTIBULAR APPARATUS

The vestibular apparatus is the organ that detects sensations concerned with equilibrium. It is composed of a system of bony tubes and chambers in the petrous portion of the temporal bone called the *bony labyrinth* and within this a system of membranous tubes and chambers called the *membranous labyrinth,* which is the functional part of the apparatus. The top of Figure 52–3 illustrates the membranous labyrinth; it is composed mainly of the *cochlear duct,* three *semicircular canals,* and two large chambers known as the *utricle* and the *saccule.* The cochlear duct is the major sensory area for hearing and has nothing to do with equilibrium. However, the *utricle,* the *semicircular canals,* and probably the *saccule* are all integral parts of the equilibrium mechanism.

The Macula—The Sensory Organ of the Utricle and the Saccule for Detecting the Orientation of the Head with Respect to Gravity. Located on the inside surface of each utricle and saccule is a small sensory area slightly over two mm in diameter called a *macula.* The macula of the utricle lies in the horizontal plane on the inferior surface of the utricle and plays an important role in determining the normal orientation of the head with respect to the direction of gravitational or *acceleratory forces.* On the other hand, the macula of the saccule is located in a vertical plane on the medial wall of the saccule. Some

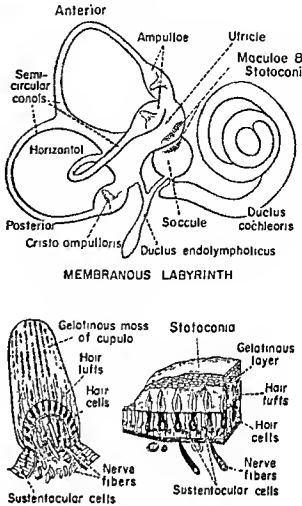


Figure 52–3. The membranous labyrinth, and organization of the crista ampullaris and the macula. (Modified from Gross Gray's Anatomy of the Human Body Lea & Febiger; modified from Kolmer by Buchanan. Functional Neuroanatomy Lea & Febiger)

research studies suggest that it operates in close association with the cochlear duct for detecting certain types of sounds and therefore may not be very important as an equilibrium apparatus. However, it is possible if not probable that the saccule also operates as an equilibrium apparatus, especially when the head is not in a vertical position.

Each macula is covered by a gelatinous layer in which many small calcium carbonate crystals called *statoconia* (or *otoliths*) are imbedded. Also, in the macula are thousands of *hair cells,* one of which is illustrated in Figure 52–4; these project *cilia* up into the gelatinous layer. The bases and sides of the hair cells synapse with sensory axons of the vestibular nerve.

Even under resting conditions, most of the nerve fibers leading from the hair cells transmit a continuous series of nerve impulses, averaging about 200 per second. Bending the cilia of a hair cell to *one side* causes the impulse traffic in its nerve fibers to increase markedly; bending the cilia to the opposite side decreases the impulse traffic,

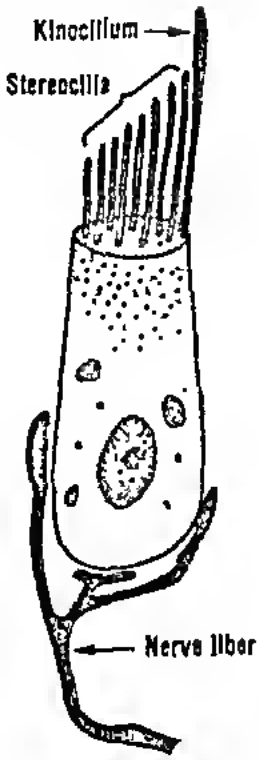


Figure 52–4. A hair cell of the membranous labyrinth of the equilibrium apparatus.

often turning it off completely. Therefore, as the orientation of the head in space changes and the weight of the otoconia (whose specific gravity is about 3 times that of the surrounding tissues) bends the cilia, appropriate signals are transmitted to the brain to control equilibrium.

In each macula the different hair cells are oriented in different directions so that some of them are stimulated when the head bends forward, some when it bends backward, others when it bends to one side, and so forth. Therefore, a different pattern of excitation occurs in the macula for each position of the head; it is this "pattern" that apprises the brain of the head's orientation.

The Semicircular Canals. The three semicircular canals in each vestibular apparatus, known respectively as the *anterior*, *posterior*, and *horizontal semicircular canals*, are arranged at right angles to each other so that they represent all three planes in space. When the head is bent forward approximately 30 degrees, the two horizontal semicircular canals are located approximately horizontal with respect to the surface of the earth. The anterior canals are then located in vertical planes that project *forward and 45 degrees outward*, and the posterior canals are also then in vertical planes but project *backward and 45 degrees outward*. Thus, the anterior canal on each side of the head is in a plane parallel to that of the posterior canal on the opposite side of the head, whereas the two horizontal canals on the two sides are located in [illegible] the same plane.

Each semicircular canal has an enlargement at one of its ends called the *ampulla*, and the canals are filled with a viscous fluid called *endolymph*. Flow of this fluid in the canals excites a sensory organ in the ampulla. As illustrated in Figure 52–3, located in each ampulla is a small crest called a *crista ampullaris*, and on top of the crista is a gelatinous mass similar to that in the utricle and known as the *cupula*. Into the cupula are projected cilia from hair cells located along the ampullary crest, and these hair cells in turn are connected to sensory nerve fibers that pass into the *vestibular nerve*. Bending the cupula to one side, caused by flow of fluid in the canal, stimulates the hair cells, while bending in the opposite direction inhibits them. Thus, appropriate signals are sent through the vestibular nerve to apprise the central nervous system of fluid movement in the respective canal.

Directional Sensitivity of the Hair Cells—The Kinocilium. As illustrated in Figure 52–4, each hair cell, whether in a macula or a cupula, has an average of about 50 small cilia, called *stereocilia*, plus one very large cilium called the *kinocilium*. This kinocilium is located to one side of the hair cell, always on the same side of the cell with respect to its orientation on the ampullary crest. This is the cause of the directional sensitivity of the hair cells: namely, stimulation when the cilia are bent toward the kinocilium side and inhibition when bent in the opposite direction.

Neuronal Connections of the Vestibular Apparatus with the Central Nervous System. Figure 52–5 illustrates the central connections of the vestibular nerve. Most of the vestibular nerve fibers end in the vestibular nuclei, which are located approximately at the junction of the medulla and the pons, but some fibers pass without synapsing to the fastigial nuclei, uvula, and flocculonodular lobes of the cerebellum. The fibers that end in the vestibular nuclei synapse with second order neurons that also send fibers into these areas of the cerebellum as well as to the cortex of other portions of the cerebellum, into the vestibulospinal tract, into

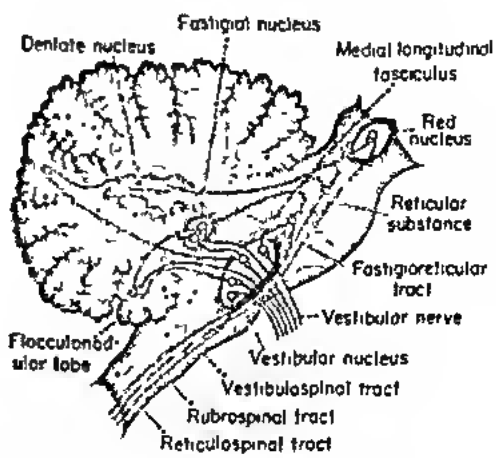


Figure 52–5. Connections of vestibular nerves in the central nervous system.

the medial longitudinal fasciculus, and to other areas of the brain stem, particularly the reticular formation.

Note especially the very close association between the vestibular apparatus, the vestibular nuclei, and the cerebellum. The primary pathway for the reflexes of equilibrium begins in the vestibular nerves and passes next to both the vestibular nuclei and the cerebellum. Then, along with two-way traffic of impulses between these two, signals are also sent into the reticular nuclei of the brain stem as well as down the spinal cord via the vestibulospinal and reticulospinal tracts. In turn, the signals to the cord control the interplay between facilitation and inhibition of the antigravity muscles, thus automatically controlling equilibrium.

The *flocculonodular lobes* of the cerebellum seem to be especially concerned with equilibrium functions of the semicircular canals because destruction of these lobes gives almost exactly the same clinical symptoms as destruction of the semicircular canals themselves. That is, severe injury to either of these structures causes loss of equilibrium during *rapid changes in direction of motion* but does not seriously disturb equilibrium under static conditions, as will be discussed in subsequent sections. It is also believed that the *uvula* of the cerebellum plays an equally important role in static equilibrium.

Signals transmitted upward in the brain stem from both the vestibular nuclei and the cerebellum via the *medial longitudinal fasciculus* cause corrective movements of the eyes every time the head rotates so that the eyes can remain fixed on a specific visual object. Signals also pass upward (either through this same tract or through reticular tracts) to the cerebral cortex, probably terminating in a primary cortical center for equilibrium located in the parietal lobe deep in the Sylvian fissure, on the opposite side of the fissure from the auditory area of the superior temporal gyrus. These signals apprise the psyche of the equilibrium status of the body.

The vestibular nuclei on either side of the brain stem are divided into four separate subdivisions. These are:

(1 and 2) The *superior* and *medial vestibular nuclei* that receive signals mainly from the semicircular canals and in turn send large numbers of nerve signals into the *medial longitudinal fasciculus* to cause corrective movements of the eyes as well as signals through the *medial vestibulospinal tract* to cause appropriate movements of the neck and head.

(3) The *lateral vestibular nucleus* that receives its innervation primarily from the utricle and possibly saccule, and in turn transmits outflow signals to the spinal cord through the *lateral vestibulospinal tract* to control body movement.

(4) The *inferior vestibular nucleus* that receives signals from both the semicircular canals and the utricle and in turn sends signals into both the cerebellum and the reticular formation of the brain stem.

FUNCTION OF THE UTRICLE (AND PROBABLY SACCULE) IN THE MAINTENANCE OF STATIC EQUILIBRIUM

It is especially important that the different hair cells are oriented in all different directions in the maculae of the utricles and saccules so that at different positions of the head, different hair cells become stimulated. The "patterns" of stimulation of the different hair cells apprise the nervous system of the position of the head with respect to the pull of gravity. In turn, the vestibular, cerebellar, and reticular motor systems reflexly excite the appropriate muscles to maintain proper equilibrium.

The maculae in the utricles function extremely effectively for maintaining equilibrium when the head is in the near-vertical position. Indeed, a person can determine as little as a half-degree of mal-equilibrium when the head leans from the precise upright position. On the other hand, as the head is leaned farther and farther from the upright, the determination of head orientation by the vestibular sense becomes poorer and poorer. Obviously, extreme sensitivity in the upright position is of major importance for maintenance of precise vertical static equilibrium, which is the most essential function of the vestibular apparatus.

Detection of Linear Acceleration by the Maculae. When the body is suddenly thrust forward—that is, when the body accelerates—the statoconia, which have greater inertia than the surrounding fluids, fall backward on the hair cell cilia, and information of mal-equilibrium is sent into the nervous centers, causing the individual to feel as though he were falling backward. This automatically causes him to lean his body forward until the anterior shift of the statoconia caused by leaning exactly equals the tendency for the statoconia to fall backward. At this point, the nervous system detects a state of proper equilibrium and therefore leans the body no farther forward. Thus, the maculae operate to maintain equilibrium during linear acceleration in exactly the same manner as they operate in static equilibrium.

The maculae *do not* operate for the detection of linear *velocity*. When runners first begins to run, they must lean far forward to keep from falling over backward because of *acceleration*, but once they have achieved running speed, they would not have to lean forward at all if they were running in a vacuum. When running in air they lean forward to maintain equilibrium only because of the air resistance against their bodies, and in this instance it is not the maculae that make them lean but the pressure of the air acting on pressure end-organs in the skin, which initiate the appropriate equilibrium adjustments to prevent falling.

THE SEMICIRCULAR CANALS AND THEIR DETECTION OF HEAD ROTATION

When the head suddenly *begins* to rotate in any direction (this is called angular acceleration), the endolymph in the membranous semicircular canals, because of its inertia, tends to remain stationary while the semicircular canals themselves turn. This causes relative fluid flow in the canals in the direction opposite to the rotation of the head.

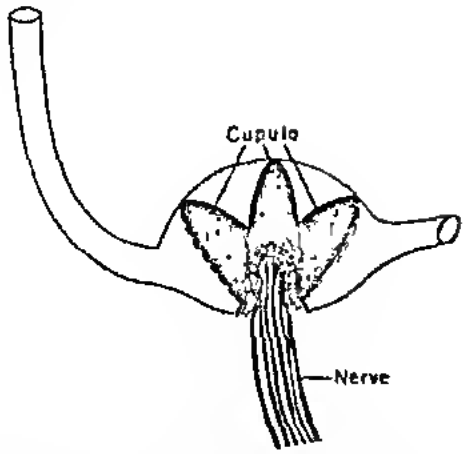


Figure 52–6. Movement of the cupula and its embedded hairs during rotation first in one direction and then in the opposite direction. (Often the top of the cupula is attached so that it merely bulges in one direction or the other.)

Figure 52–6 illustrates an *ampulla* of one of the semicircular canals, showing the *cupula* and its embedded hairs bending in the direction of fluid movement. And Figure 52–7 illustrates the discharge signal from a single hair cell in the crista ampullaris when an animal is rotated for 40 seconds, showing that (1) even when the cupula is in its resting position the hair cell emits a tonic discharge of approximately 200 impulses per second, (2) when the animal is rotated, the hairs bend to one side and the rate of discharge increases greatly; and (3) with continued rotation, the excess discharge of the hair cell gradually subsides back to the resting level in about 20 seconds.

The reason for this adaptation of the receptor is that within a second or more of rotation, friction in the semicircular canal causes the endolymph to rotate as rapidly as the semicircular canal itself; then in an additional 15 to 20 seconds the cupula slowly returns to its resting position in the middle of the ampulla because of its own elastic recoil.

When the rotation suddenly stops, exactly the opposite effects take place: the endolymph continues to rotate while the semicircular canal stops. This time the cupula is bent in the opposite direction, causing the hair cell to stop discharging entirely. After another few seconds, the endolymph stops moving, and the cupula returns gradually to its resting position in about 20 seconds, thus allowing the discharge of the hair cell to return to its normal tonic level as shown to the right in Figure 52–7.

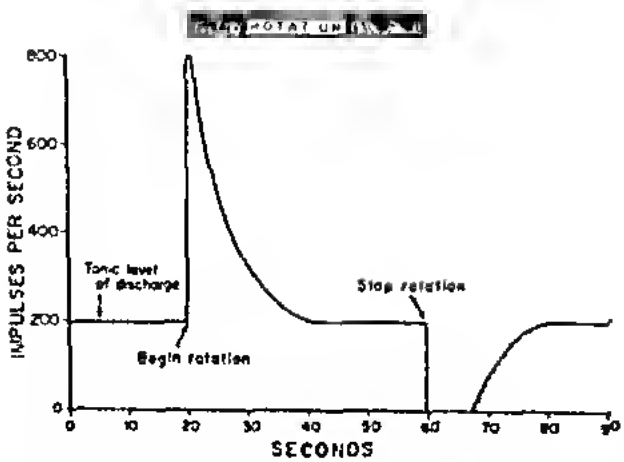


Figure 52–7. Response of a hair cell when a semicircular canal is stimulated first by rotation and then by stopping rotation.

Thus, the semicircular canal transmits a positive signal when the head *begins* to rotate and a negative signal when it *stops* rotating. Furthermore, at least some hair cells will always respond to rotation in any plane—horizontal, sagittal, or coronal—for fluid movement always occurs in at least one semicircular canal.

Rate of Angular Acceleration Required to Stimulate the Semicircular Canals. The angular acceleration required to stimulate the semicircular canals in the human being averages about 1 degree per second per second. In other words, when one begins to rotate, the velocity of rotation must be as much as 1 degree per second by the end of the first second, 2 degrees per second by the end of the second second, 3 degrees per second by the end of the third second, and so forth, in order for the person barely to detect that the rate of rotation is increasing.

"Predictive" Function of the Semicircular Canals in the Maintenance of Equilibrium. Since the semicircular canals do not detect that the body is off balance in the forward direction, in the side direction, or in the backward direction, one might at first ask: What is the function of the semicircular canals in the maintenance of equilibrium? All they detect is that the person's head is beginning to rotate or stopping rotation in one direction or another. Therefore, the function of the semicircular canals is not likely to be to maintain static equilibrium or to maintain equilibrium during linear acceleration or when the person is exposed to steady centrifugal forces. Yet loss of function of the semicircular canals causes a person to have very poor equilibrium when attempting to perform *rapid* and *intricate* body movements.

We can explain the function of the semicircular canals best by the following illustration. If a person is running forward rapidly, and then suddenly begins to turn to one side, he falls off balance a second or so later unless appropriate corrections are made *ahead of time*. But, unfortunately, the macula of the utricle cannot detect that he is off balance until *after* this has occurred. On the other hand, the semicircular canals will have already detected that the person is turning, and this information can easily apprise the central nervous system of the fact that the person *will* fall off balance within the next second or so unless some correction is made. In other words, the semicircular canal mechanism *predicts ahead of time* that mal-equilibrium is going to occur even before it does occur and thereby causes the equilibrium centers to make appropriate preventive adjust-

ments. In this way, the person need not fall off balance before he begins to correct the situation.

Removal of the flocculonodular lobes of the cerebellum prevents normal function of the semicircular canals but does not prevent normal function of the macular receptors. It is especially interesting in this connection that the cerebellum serves as a "predictive" organ for most of the other rapid movements of the body as well as those having to do with equilibrium. These other functions of the cerebellum are discussed in the following chapter.

VESTIBULAR POSTURAL REFLEXES

Sudden changes in the orientation of an animal in space elicit reflexes that help to maintain equilibrium and posture. For instance, if an animal is suddenly pushed to the right, even before it can fall more than a few degrees its right legs extend instantaneously. In other words, this mechanism *anticipates* that the animal will be off balance in a few seconds and makes appropriate adjustments to prevent this.

Another type of vestibular postural reflex occurs when the animal suddenly falls forward. When this occurs, the forepaws extend forward, the extensor muscles tighten, and the muscles in the back of the neck stiffen to prevent the animal's head from striking the ground. This reflex is probably also of importance in locomotion, for, in the case of the galloping horse, the downward thrust of the head can automatically provide reflex thrust of the forelimbs to move the animal forward for the next gallop.

VESTIBULAR MECHANISM FOR STABILIZING THE EYES AND FOR NYSTAGMUS

When a person changes his direction of movement rapidly, or even leans his head sideways, forward, or backward, it would be impossible for him to maintain a *stable image on the retinae of his eyes unless he had* some automatic control mechanism to stabilize the direction of gaze of the eyes. In addition, the eyes would be of little use in detecting an image unless they remained "fixed" on each object long enough to gain a clear image. Fortunately, each time the head is suddenly rotated, signals from the semicircular canals cause the eyes to rotate in an equal and opposite direction to the rotation of the head. This results from reflexes transmitted from the canals through the *vestibular nuclei* and the *medial longitudinal fasciculus* to the *ocular nuclei.*

OTHER FACTORS CONCERNED WITH EQUILIBRIUM

The Neck Proprioceptors. The vestibular apparatus detects the orientation and movements *only of the head.* Therefore, it is essential that the nervous centers also receive appropriate information depicting the orientation of the head with respect to the body. This information is transmitted from the proprioceptors of the neck and body directly into the vestibular and reticular nuclei of the brain stem and also indirectly by way of the cerebellum.

By far the most important proprioceptive information needed for the maintenance of equilibrium is that derived from the *joint receptors of the neck.* When the head is leaned in one direction by bending the neck, impulses from the neck proprioceptors keep the vestibular apparatuses from giving the person a sense of mal-equilibrium. They do this by transmitting signals that exactly oppose the signals transmitted from the vestibular apparatuses. However, *when the entire body* leans in one direction, the impulses from the vestibular apparatuses *are not opposed* by the neck proprioceptors; therefore, the person in this instance does perceive a change in equilibrium status.

The Neck Reflexes. In an animal *whose vestibular apparatuses have been destroyed,* bending the neck causes immediate muscular reflexes called *neck reflexes* occurring especially in the forelimbs. For instance, bending the head forward causes both forelimbs to relax. However, when the vestibular apparatuses are *intact, this effect does not occur because the vestibular reflexes* function exactly oppositely to the neck reflexes. Thus, if the head is flexed downward, the vestibular reflex tends to extend the forelimbs, while the neck reflexes tend to relax them. Since the equilibrium of the entire body and not of the head alone must be maintained, it is easy to understand that the vestibular and neck reflexes must function oppositely. Otherwise, each time the neck should bend, the animal would immediately fall off balance.

Proprioceptive and Exteroceptive Information from Other Parts of the Body. Proprioceptive information from other parts of the body besides the neck is also important in the maintenance of equilibrium. For instance, pressure sensations from the footpads can tell one (a) whether weight is distributed equally between the two feet and (b) whether weight is more forward or backward on the feet.

An instance in which exteroceptive information is necessary for maintenance of equilibrium occurs when a person is running. The air pressure against the front of the body signals that a force is opposing the body in a direction different from that caused by gravitational pull; as a result, the person leans forward to oppose this.

Importance of Visual Information in the Maintenance of Equilibrium. After complete destruction of the vestibular apparatuses, and even after loss of most proprioceptive information from the body, a person can still use the visual mechanisms effectively for maintaining equilibrium. Even slight linear or rotational movement of the body instantaneously shifts the visual images on the retina, and this information is relayed to the equilibrium centers. Many persons with complete destruction of the vestibular apparatuses have almost normal equilibrium as long as their eyes are open and as long as they perform all motions slowly. But, when moving rapidly or when the eyes are closed, equilibrium is immediately lost.

FUNCTIONS OF THE RETICULAR FORMATION AND SPECIFIC BRAIN STEM NUCLEI IN CONTROLLING SUBCONSCIOUS, STEREOTYPED MOVEMENTS

Rarely, a child called an *anencephalic monster* is born without brain structures above the mesencephalic region, and some of these children have been kept alive for many months. They are able to perform essentially

all the functions of feeding, such as sucking, extrusion of unpleasant food from the mouth, and moving the hands to the mouth to suck the fingers. In addition, they can yawn and stretch. They can cry and follow objects with the eyes and by movements of the head. Also, placing pressure on the upper anterior parts of their legs will cause them to pull to the sitting position.

Therefore, it is obvious that many of the stereotyped motor functions of the human being are integrated in the brain stem. Unfortunately, the loci of most of the different motor control systems have not been found except for the following.

Stereotyped Body Movements. Most movements of the trunk and head can be classified into several simple movements, such as forward flexion, extension, rotation, and turning movements of the entire body. These types of movements are controlled by special nuclei located mainly in the mesencephalic and lower diencephalic region. For instance, *rotational movements* of the head and eyes are controlled by the *interstitial nucleus*, which is illustrated in Figure 52–1. This nucleus lies in close approximation to the *medial longitudinal fasciculus*, through which it transmits a major portion of its control impulses. The *raising movements* of the head and body are controlled by the *prestitial nucleus*, which is located approximately at the juncture of the diencephalon and mesencephalon. On the other hand, the *flexing movements* of the head and body are controlled by the *nucleus precommissuralis* located at the level of the posterior commissure. Finally, the *turning movements* of the entire body, which are much more complicated, involve both the pontile and mesencephalic reticular formation.

Function of the Subthalamic Areas—Forward Progression. Much less is known about function of higher brain stem centers in posture and locomotion than of the lower centers, principally because of the complexity of the neuronal connections. However, it is known that stimulation of centers in or around the subthalamic nuclei can cause rhythmic limb motions, including forward walking reflexes. This does not mean that the individual muscles of walking are controlled from this region but simply that excitation of this region sends "command" signals to activate the cord centers where *preprogrammed* neuronal circuits then cause the actual walking movements.

A cat with its brain transected beneath the thalamus but above the subthalamus can walk in an almost completely normal fashion—so much so that the observer often cannot tell the difference. However, when the animal comes to an obstruction it simply butts its head against the obstruction and tries to keep on walking. Thus, it lacks *purposefulness of locomotion*.

The function of the subthalamic region in walking is frequently described as that of controlling *forward progression*.

THE BASAL GANGLIA—THEIR MOTOR FUNCTIONS

Physiologically, the basal ganglia are considered to be comprised of the *caudate nucleus, putamen*, and *globus pallidus*. However, the *substantia nigra, subthalamus*, and important portions of both the *thalamus* and *reticular formation* operate in association with these and therefore are actually part of the *basal ganglia system* for motor

The basal ganglia function in close association with the cerebral cortex for both motor control and sensory activities. However, little is known about the sensory functions of the basal ganglia despite the fact that as many nervous pathways connect the sensory areas of the cortex with the basal ganglia as with the motor areas. One of the most important recent findings has been that, when the body performs a muscle activity, action potentials appear in the basal ganglia before they appear in the cortical motor areas. And it is very probable that the basal ganglia themselves are most often set into activity not by signals from the motor cortex but instead by signals directly from the sensory and sensory association portions of the cortex, or from the thalamus and lower brain stem. Thus, there is much belief that the basal ganglia play an essential role in the initiation of most if not all motor activities of the body besides some of the involuntary muscle contractions that originate in the brain stem or spinal cord.

Even though the basal ganglia are closely associated with the cerebral cortex in normal motor control, the basal ganglia can still provide many aspects of motor control even after the cerebral cortex has been removed. However, the motor activities are then much grosser and less determinative than when the cortex contributes its capabilities as well. We shall say more about the interrelationships between the basal ganglia and the motor cortex in the following chapter.

Physiologic Anatomy of the Basal Ganglial Neuronal Connections

Figure 52–8 illustrates the complexity of the interconnections among the different basal ganglia as well as between the basal ganglia and the brain stem, the

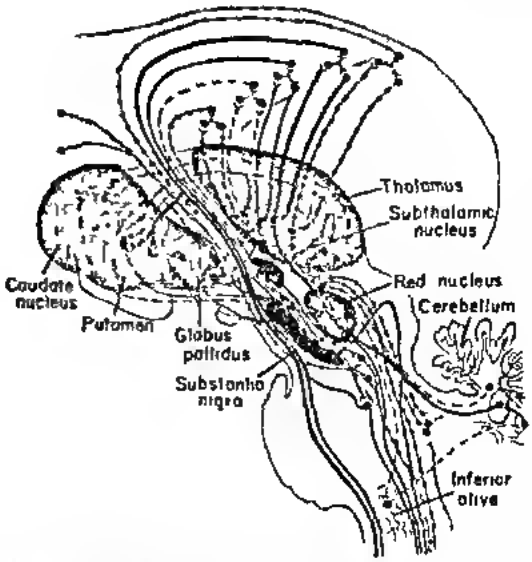


Figure 52–8. Pathways through the basal ganglia and related structures of the brain stem, thalamus, and cerebral cortex. (From Jung and Hassler: Handbook of Physiology, Sec. I, Vol. II, Baltimore, Williams & Wilkins Company, 1960.)

thalamus, and the cerebral cortex. However, several partially distinct neuronal pathways have been discerned among these multiple connectivities, as follows:

(1) Pathway from Cortex to Basal Ganglia and Back to Cortex. Figure 52–9 illustrates the pathway that involves the greatest number of nerve fibers in the basal ganglia system. It begins in all areas of the cerebral cortex and passes through the following four stages: (a) From the *cortex* to the *caudate nucleus* and *putamen* (which together are called the *neostriatum*, or sometimes the *striate body*). (b) From the neostriatum (caudate nucleus and putamen) to the *globus pallidus*, and eventually entering the medial segment of the globus pallidus as illustrated in the figure. (c) From the medial segment of the globus pallidus to the *ventroanterior* and *ventrolateral nuclei of the thalamus*. It is to these same thalamic nuclei that signals from the cerebellum are also transmitted. (d) From the ventroanterior and the ventrolateral nuclei of the thalamus *back to all areas of the cortex*.

Essentially all the nerve fibers from the neostriatum to the globus pallidus (stage b in the above circuit) are inhibitory pathways, secreting the inhibitory transmitter *GABA* at their nerve endings. *Because of this inhibitory step, this circular pathway beginning in the cortex and returning once again to the cortex is a negative feedback loop* that provides stability to many aspects of the motor control interactions.

(2) Mutual Inhibitory Pathways Between the Neostriatum and the Substantia Nigra. Figure 52–10 illustrates a pathway from the *caudate nucleus* and *putamen* (the neostriatum) to the *substantia nigra* that secretes the inhibitory transmitter *GABA* at its terminus in the substantia nigra. In turn, a group of dark-staining cells in the substantia nigra sends axons back to the caudate nucleus and putamen, secreting the inhibitory transmitter *dopamine* at their terminations. This mutual inhibitory pathway normally maintains a

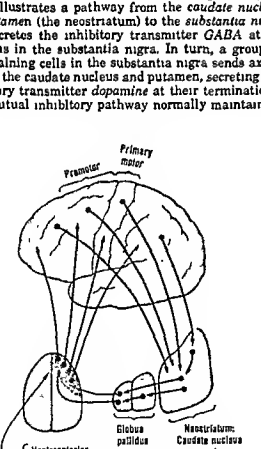


Figure 52–9. Feedback circuit from the cerebral cortex to the basal ganglia, then through the thalamus back to the cortex.

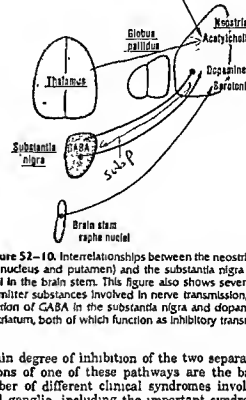


Figure 52–10. Interrelationships between the neostriatum (caudate nucleus and putamen) and the substantia nigra and raphe nuclei in the brain stem. This figure also shows several different transmitter substances involved in nerve transmission, especially secretion of GABA in the substantia nigra and dopamine in the neostriatum, both of which function as inhibitory transmitters.

certain degree of inhibition of the two separate areas. Lesions of one of these pathways are the basis of a number of different clinical syndromes involving the basal ganglia, including the important syndrome Parkinson's disease that we shall discuss later.

Figure 52–10 also shows that nerve fibers from the cerebral cortex (as well as others from the thalamus) secrete *acetylcholine* in the caudate nucleus and putamen, causing excitatory effects. And still another pathway, from the brain stem raphe nuclei, secretes *serotonin*. Serotonin is an inhibitory transmitter; this pathway is believed to function primarily in relation to the sleep-producing effects of the raphe nuclei.

And not shown in this figure is still another tract that passes directly from the caudate nucleus to the substantia nigra that secretes *substance P*, an excitatory transmitter, illustrating even more the complexity of the different neurotransmitters involved in the interactions within the basal ganglia system.

(3) Interrelationships Between the Thalamic Intralaminar Nuclei and the Basal Ganglia. Figure 52–11 illustrates a feedback loop between the intralaminar nuclei of the thalamus and the basal ganglia that is similar to the feedback loop between the cerebral cortex and the basal ganglia. Note also in the figure the close relationship of the intralaminar nuclei with the lower brain stem. These together are older portions of the nervous system and presumably are important in the control of the more primitive aspects of motor control.

(4) Outflow Tracts from the Basal Ganglia Toward the Lower Brain Stem. Even though it is known that the basal ganglia can have important effects on the motor control functions of the brain stem, there are surprisingly few direct nerve fiber projections from the basal ganglia to the lower brain stem. Nevertheless, shown in Figure 52–12 are several pathways, all of which are relatively small, that pass from the basal ganglia to the brain stem reticular formation. Most of the signals pass from the globus pallidus through either the subthalamic nucleus or the substantia nigra before entering the reticular formation. However, a few fibers

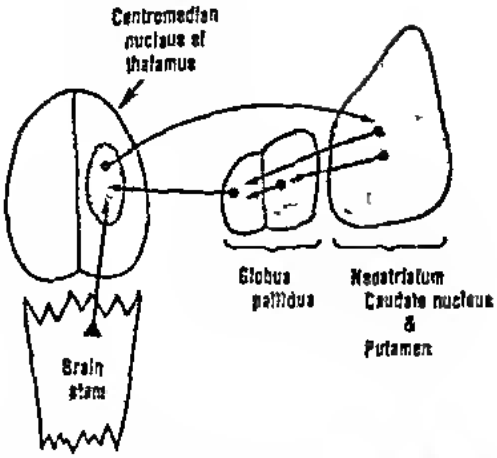


Figure 52–11. An older pathway involving the basal ganglia, showing interplay of signals among the basal ganglia, the centromedian nucleus of the thalamus, and the brain stem

pass from the medial portion of the globus pallidus directly to the brain stem

FUNCTIONS OF THE BASAL GANGLIA

Before attempting to discuss the functions of the basal ganglia in human beings, we should speak briefly of the better known functions of these ganglia in lower animals. In birds, for instance, the cerebral cortex is poorly developed while the basal ganglia are highly developed. These ganglia perform essentially all the motor functions, even controlling the voluntary movements in much the same manner that the motor cortex of the human being controls voluntary movements. Furthermore, in the cat, and to a lesser extent in the dog, decortication removes only the discrete types of motor functions and does not interfere with the animal's ability to walk, eat, fight, develop rage, have periodic sleep and wakefulness, and even

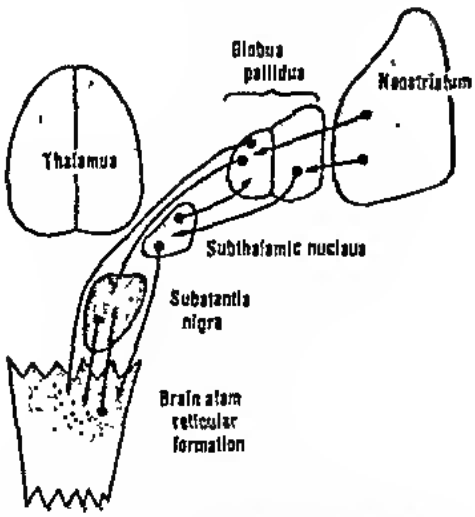


Figure 52–12. Pathway for transmission of signals from the basal ganglia to the brain stem reticular formation by way of the subthalamic nucleus and the substantia nigra.

participate naturally in sexual activities. However, if a major portion of the basal ganglia is destroyed, only gross stereotyped movements remain, which were discussed earlier in relation to the mesencephalic animal.

Finally, in the human being, cortical lesions in very young individuals destroy the discrete movements of the body, particularly of the hands and distal portions of the lower limbs, but do not destroy the person's ability to walk crudely, to control equilibrium, or to perform many other subconscious types of movements. However, simultaneous destruction of a major portion of the caudate nucleus almost totally paralyzes the opposite side of the body except for a few stereotyped reflexes integrated in the cord or brain stem.

With this brief background of the overall function of the basal ganglia, we can attempt to dissect the functions of certain portions of the basal ganglia system, realizing that the system actually operates, along with the motor cortex and cerebellum, as a total unit and that individual functions cannot be ascribed to the different individual parts of the basal ganglia.

Inhibition of Motor Tone by the Basal Ganglia. Though it is wrong to ascribe a single function to all the basal ganglia, nevertheless, one of the general effects of diffuse basal ganglia excitation is to inhibit muscle tone throughout the body. This effect results from inhibitory signals transmitted from the basal ganglia to both the motor cortex and the lower brain stem. Therefore, whenever widespread destruction of the basal ganglia occurs, this causes muscle rigidity throughout the body. For instance, when the brain stem is transected at the mesencephalic level, which removes the inhibitory effects of the basal ganglia, the phenomenon of decerebrate rigidity occurs.

Yet, despite this general inhibitory effect of the basal ganglia, stimulation of certain specific areas within the basal ganglia can elicit positive muscle contractions and at times even complex patterns of movements.

Function of the Caudate Nucleus and Putamen (The Neostriatum). The caudate nucleus and putamen seem to function together to initiate and regulate gross intentional movements of the body. To perform this function they transmit impulses through two different pathways: (1) into the *globus pallidus*, thence by way of the *thalamus* to the cerebral cortex, and finally downward into the spinal cord through the *corticospinal pathway;* (2) downward through the *globus pallidus* and the *substantia nigra* by way of short axons into the *reticular formation* and finally into the spinal cord mainly through the *reticulospinal tracts*.

In summary, the neostriatum helps control gross intentional movements that we normally perform subconsciously. However, this control also involves the motor cortex, with which the neostriatum is closely connected.

Function of the Globus Pallidus. It is already

clear that almost all the outflow of signals from the basal ganglia are channelled through the globus pallidus en route back to the cortex or on their way to lower brain centers. However, in addition to this motor relay function of the globus pallidus, the globus pallidus seems to have still another function that operates in close association with the subthalamus and brain stem to help control the axial and girdle movements of the body. These movements provide the background positioning of the body and proximal limbs so that the more discrete motor functions of the hands and feet can then be performed. That is, a person wishing to perform an exact function with a hand first positions the body, next positions the legs and arms, and finally tenses all the axial and girdle muscles to provide background positioning and stability of all the proximal portions of the body. These associated tonic contractions are supposedly initiated by circuits in the globus pallidus but operate also through the axial and girdle motor control areas of the brain stem. Lesions of the globus pallidus seriously interfere with the attitudinal movements that are necessary to position the hand and, therefore, make it difficult or impossible for one to use the hand for discrete activities.

Electrical stimulation of the globus pallidus while an animal is performing a gross body movement often will stop the movement in a static position, the animal holding that position for many seconds while the stimulation continues. This fits with the concept that the globus pallidus is involved in some type of servo feedback motor control system that is capable of locking the different parts of the body into specific positions.

CLINICAL SYNDROMES RESULTING FROM DAMAGE TO THE BASAL GANGLIA

Much of what we know about the function of the basal ganglia comes from study of patients with basal ganglia lesions whose brains have undergone pathologic studies after death. Among the different clinical syndromes are:

Chorea. Chorea is a disease in which random uncontrolled "flicking-type" movements occur one after another. Normal progression of movements cannot occur; instead, the person may perform one pattern of movement for a few seconds and then suddenly another.

A particularly important type of chorea is *Huntington's chorea*, which is a hereditary disorder that begins manifesting usually in the third or fourth decade of life. It is characterized at first by flicking movements but then by progressively severer distortional movements of the body as the disorder becomes more severe. In addition, severe dementia develops concurrently with the motor dysfunctions.

In Huntington's chorea there is diffuse and widespread degeneration of neurons in both the caudate nucleus and the putamen. The neurons most involved are those that secrete the inhibitory transmitter GABA, but some of the acetylcholine-secreting neurons also degenerate. Failure of the GABA-secreting neurons removes the inhibition of the striatonigral pathway and allows the dopamine-secreting neurons of the substantia nigra to become greatly overactive. This in turn allows excessive secretion of dopamine by the nigrostriatal fibers back in the caudate nucleus and putamen, which greatly inhibits both these nuclei. In addition, degeneration of the acetylcholine-secreting neurons adds further to the depression of these nuclei. Presumably, this depressed function interrupts some of the feedback loops from the cortex to the basal ganglia and then back to the cortex through the thalamus, thus causing the abnormal movements.

Thus, Huntington's chorea is believed to be caused by widespread depression of function in the caudate nucleus and putamen. If we also remember that at least half of all the cortical connections of the caudate nucleus and putamen are with sensory areas of the cortex rather than with motor areas, this could also explain the generalized dementia that occurs in Huntington's chorea.

Athetosis. In this disease, slow, writhing movements of a hand, the neck, the face, the tongue, or some other part of the body occur continually. The movements are likely to be wormlike, first with overextension of the hands and fingers, then flexion, then rotary twisting to the side—all these continuing in a slow, rhythmic, repetitive writhing pattern. The contracting muscles exhibit a high degree of spasm, and the movements are enhanced by emotions or by excessive signals from the sensory organs. Furthermore, voluntary movements in the affected area are greatly impaired or sometimes even impossible.

The damage in athetosis is usually found in the *lateral portion of the globus pallidus* or in this area and the *neostriatum* as well. Athetosis is usually attributed to the interruption of feedback circuits among the basal ganglia, thalamus, and cerebral cortex. The normal feedback circuits presumably allow a constant and rapid interplay between antagonistic muscle groups so that finely controlled static or progressive movements can take place. However, if the feedback circuits are blocked, it is supposed that the detouring impulses may take devious routes through the basal ganglia, thalamus, and motor cortex, causing a succession of abnormal movements.

Hemiballismus. Hemiballismus is an uncontrollable succession of violent movements of large areas of the body. These may occur once every few seconds or sometimes only once in many minutes. For instance, an entire leg might suddenly jerk uncontrollably to full flexion, or the entire trunk might go through an extreme, sudden torsion movement, or an arm might be pulled upward suddenly with great force. Hemiballismus of the legs or trunk causes the person to fall to the ground if walking, and even in bed, the person tosses violently when affected by these powerful and strong intermittent movements. Furthermore, attempts to perform voluntary movements frequently invoke ballistic movements in place of the normal movements.

Hemiballismus on one side of the body results from a *large lesion in the opposite subthalamus.* The smooth, progressive or rhythmic movements of the limbs or other parts of the body normally integrated in this area can no longer occur, but excitatory impulses attempting to evoke such movements elicit instead uncontrollable ballistic movements.

Parkinson's Disease. Parkinson's disease, which is also known as *paralysis agitans*, results almost invariably from *widespread destruction of the substantia nigra* but is often associated also with lesions of the *globus*

pallidus and other related areas. It is characterized by (1) *rigidity* of the musculature either in widespread areas of the body or in isolated areas, (2) *tremor at rest* of the involved areas in most but not all instances, and (3) a serious *inability to initiate movement*, called *akinesia*.

The cause or causes of these abnormal motor activities are almost entirely unknown. However, they are almost certainly related to *loss of dopamine secretion* in the putamen and caudate nucleus by the nerve endings of the *nigrostriatal tract*. Destruction of the substantia nigra causes this tract to degenerate and the dopamine normally secreted in the putamen and caudate nucleus no longer to be present. But still present are large numbers of neurons that secrete acetylcholine, and these transmit excitatory signals throughout the basal ganglia. It is also believed that the dopamine from the nigrostriatal pathway normally acts to inhibit these acetylcholine-producing neurons or in some other way to counter their activity. But, in the absence of dopamine secretion, the acetylcholine pathways become overly active, which presumably is the basis for many of the motor symptoms in Parkinson's disease.

The rigidity in this disease is somewhat different from that which occurs in decerebrate rigidity, for decerebrate rigidity results mainly from hyperactivity of the muscle spindle system, and even slight movement of a muscle is met with rather extreme reflex resistance resulting from feedback through the stretch reflex mechanism. Parkinsonian rigidity, on the other hand, is more of a "plastic" type. That is, sudden movement usually is not met by intense resistance from the stretch reflexes, as in the decerebrate type of rigidity, but instead both protagonist and antagonist muscles remain tightly contracted throughout the movement. Therefore, it is believed that the rigidity of Parkinson's disease results to a great extent from excess impulses transmitted in the corticospinal system, thus activating the alpha motor fibers to the muscles, in addition to probably excess activation of the gamma efferent system as well.

Tremor usually, but not always, occurs in Parkinson's disease. Its frequency is normally four to six cycles per second. When the parkinsonian patient performs voluntary movements, the tremor becomes temporarily or partially blocked, presumably because other motor control signals—perhaps from the cerebral cortex and cerebellum—override the abnormal basal ganglial signals.

The mechanism of the tremor in Parkinson's disease is not known. However, a type of treatment that frequently relieves the tremor is surgical destruction of the *ventrolateral nucleus of the thalamus* which is one of the feedback pathways from the basal ganglia to the motor cortex. Therefore, it is presumed that loss of the inhibitory influence of dopamine in the basal ganglia leads to enhanced activity of the corticobasal ganglial-thalamic-cortical feedback circuit, leading to an oscillation that produces the muscle tremor.

Though the muscle rigidity and the tremor are both distressing to the parkinsonian patient, even more serious is the *akinesia* that occurs in the final stages of the disease. To perform even the simplest of movement, the person must exert the highest degree of concentration, and the mental effort, even mental anguish, that is necessary to make the movement "go" is often almost beyond the patient's willpower. Then, when the movement does occur, it is stiff and often staccato in character instead of occurring with smooth progression. For instance, when a patient begins to perform a discret voluntacy movement with the hands, the automat "associated" adjustments of the trunk of the body an the upper arm segments do not occur. Instead, th patient must voluntarily adjust these segments befor use of the hands. Furthermore, a tremendous amount c nervous effort must be made by the voluntary moto control system to overcome "motor stiffness" of th musculature. Thus, the person with Parkinson's diseas has a masklike face, showing almost no automatic emo tional facial expressions, he or she is usually ben forward because of the muscle rigidity; and all move ments of necessity are highly deliberate rather than th many casual subconscious movements that are normall a part of our everyday life.

The cause of the akinesia in Parkinson's disease i not known, and again we must resort to theory. It i presumed that *loss of dopamine secretion in the caudat nucleus and putamen by the nigrostriatal fibers* allow excessive activity of the acetylcholine-producing neu rons. But normal operation of the basal ganglia require a balance between both excitatory and inhibitory acti ities, and loss of this balance, in effect, leads to functionless basal ganglia system. We have alread pointed out that the basal ganglia are responsible fo many of the subconscious stereotyped movements of th body, and also responsible even for background move ments of the trunk, legs, neck, and upper arms that ar required preliminary to performing the more discret movements of the hands. If the subconscious and th background movements cannot occur, then other neur mechanisms must be substituted, especially those of th motor cortex and cerebellum. Unfortunately, though these cannot replace the movements normally controlle by the basal ganglia and certainly cannot function at subconscious level.

Treatment with L-Dopa and Anticholinergic Drugs. Administration of L-dopa to patients with Par kinson's disease ameliorates many of the symptoms especially the rigidity and the akinesia, in about tw thirds of the patients. The reason for this seems to b the following. The dopamine secreted in the caudat nucleus and putamen by the nigrostriate fibers is derivative of L-dopa. When the substantia nigra is de stroyed and the person develops Parkinson's disease, th administered L-dopa is believed to substitute for th dopamine no longer secreted by the destroyed neuron This causes more or less normal inhibition of the bas ganglia and relieves much or most of the akinesia an rigidity.

The same basal ganglia that are inhibited by dopa mine are excited by acetylcholine-secreting neurons, a was discussed earlier. Therefore, as would be expected administration of anticholinergic drugs such as scopol amine can also decrease the level of activity in the bas ganglia and therefore benefit some parkinsonian pa tients.

Coagulation of the Ventrolateral and Ventroan terior Nuclei of the Thalamus for Treatment o Parkinson's Disease. Neurosurgeons have treated Par kinson's disease patients, with varying success, by de stroying portions of the basal ganglia, thalamus, or eve motor cortex. The most prevalent treatment has bee destruction of the ventrolateral and ventroanterior nu clei of the thalamus, usually by electrocoagulation. Mos fiber pathways from the basal ganglia and cerebellum to the cerebral cortex pass through these nuclei so tha

their destruction blocks many or most of the corticobasal ganglial-thalamic-cortical feedback circuits. It is presumed that blockage of these feedbacks removes the neuronal loops that cause the tremor of Parkinson's disease.

REFERENCES

Baldessarini, R. J., and Tarsey, D. Dopamine and the pathophysiology of dyskinesias induced by antipsychotic drugs *Annu Rev Neurosci*, *3* 23, 1980

Bird, E. D. Chemical pathology of Huntington's disease. *Annu Rev Pharmacol Toxicol*, *20* 533, 1980

Bizzi, E.: The coordination of eye-head movements *Sci Am*, *231*(4) 100, 1974.

Bobath, B., and Bobath, K. Motor Development in the Different Types of Cerebral Palsy London, W Heinemann Medical Books, 1975

Brodal, A. (ed.), Basic Aspects of Central Vestibular Mechanisms New York, American Elsevier Publishing Co., 1971

Carpenter, M. B., Anatomy of the corpus striatum and brain stem *integrating systems In Brooks, V B (ed.) Handbook of Physiology* Sec. 1, Vol II Bethesda, American Physiological Society, 1981, p 947.

Chase, T. N., *et al* (eds.) Huntington's Disease New York, Raven Press, 1979

Dampney, R. A., *et al* Identification of cardiovascular cell groups in the brain stem. *Clin Exp Hypertent*, *6* 205, 1984

DeLong, M., and Georgopoulos, A. P Motor functions of the basal ganglia. *In* Brooks, V B (ed.) Handbook of Physiology Sec 1, Vol II. Bethesda, American Physiological Society, 1981, p 1017

Denny-Brown, D The Basal Ganglia. Their Relation to Disorders of Movement. New York, Oxford University Press, 1962

Di Chiara, G (ed.) GABA and the Basal Ganglia New York, Raven Press, 1981.

Divac, I., and Oberg, R. G E (eds.) The Neostriatum New York, Pergamon Press, 1979

Dublin, W. B: Fundamentals of Vestibular Pathology St Louis, Warren H. Green, Inc., 1985

Duvoisin, R. C., Parkinson's Disease New York, Raven Press, 1978

Elder, H. Y., and Trueman, E R (eds.) Aspects of Animal Movement New York, Cambridge University Press, 1980

Evarts, E. V. Brain mechanisms of movement. *Sci Am.* *241*(3) 184, 1979

Fernstrom, J. D: Role of precursor availability on control of monoamine biosynthesis in brain *Physiol Rev*, *63* 484, 1983

Galanter, E. Detection and discrimination of environmental change *In* Darian Smith, I (ed.) Handbook of Physiology Sec 1, Vol III Bethesda, American Physiological Society, 1984, p. 103

Goldberg, J M., and Fernandez, C Vestibular mechanisms *Annu Rev Physiol*, *37* 129, 1975

Goldberg, J M., and Fernandez, C: The vestibular system *In* Darian-Smith, I (ed.), Handbook of Physiology Sec. 1, Vol III Bethesda, American Physiological Society, 1984, p 977

Grant, R., and Pompeiano, O (eds.) Reflex Control of Posture and Movement. New York, Elsevier Scientific Publishing Co, 1979

Grillner, S: Control of locomotion in bipeds, tetrapods, and fish *In* Brooks, V B (ed.), Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 1179

Hobson, J A., and Brazier, M A. B (eds.) The Reticular Formation Revisited. Specifying Function for a Nonspecific System New York, Raven Press, 1980

Hood, J. D (ed.) Vestibular Mechanisms in Health and Disease New York, Academic Press, 1978

Kitai, S T.. Electrophysiology of the corpus striatum and brain stem integrating systems. *In* Brooks, V B (ed.) Handbook of Physiology. Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 997

Kuypers, H G, J, M Anatomy of the descending pathways *In* Brooks, V B (ed.), Handbook of Physiology Sec, 1, Vol, II Bethesda, American Physiological Society, 1981, p 597

Luschei, E S, and Goldberg, L J Neural mechanisms of mandibular control Mastication and voluntary biting *In* Brooks, V. B (ed) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 1237

Nauta, W J H, and Feirtag, M The organization of the brain *Sci Am*, *241*(3) 88, 1979

Olsen, R. W Drug interactions at the GABA receptor ionophore complex *Annu Rev Pharmacol Toxicol*, *22* 245, 1982

Oosterveld W J (ed) Audio-Vestibular System and Facial Nerve New York, S Karger, 1977.

Orlovsky, G N, and Shik, M L. Control of locomotion A neurophysiological analysis of the cat locomotor system *Int Rev Physiol*, *10* 281 1976

Pearson, K The control of walking *Sci Am*, *235*(6) 72, 1976

Penney, J B, Jr., and Young, A. B Speculations on the functional anatomy of basal ganglia disorders *Annu Rev Neurosci*, *6* 73, 1983

Peterson, B W Reticulospinal projections of spinal motor nuclei *Annu Rev Physiol*, *41* 127, 1979

Precht, W Vestibular mechanisms *Annu Rev Neurosci*, *2* 265, 1979

Rack, P M H Limitations of somatosensory feedback in control of posture and movement *In* Brooks, V. B (ed.) Handbook of Physiology Sec 1 Vol II Bethesda, American Physiological Society, 1981, p 229

Riklan, M L-Dopa and Parkinsonism. Springfield, Ill, Charles C Thomas, 1973

Sarno, J E., and Sarno, M T Stroke, The Condition and the Patient. New York, McGraw-Hill, 1979

Scheibel, A B The brain stem reticular core and sensory function *In* Darian Smith, I (ed) Handbook of Physiology Sec 1, Vol III Bethesda, American Physiological Society, 1984, p 213

Sherrington, C S Decerebrate rigidity and reflex coordination of movements *J Physiol (Lond.)*, *22* 319, 1898

Shik, M L., and Orlovsky, G N Neurophysiology of locomotor automatism *Physiol Rev*, *56* 465, 1976

Silverman, A J Magnocellular neurosecretory system. *Annu Rev Neurosci*, *6* 357, 1983

Stein, P S G Motor systems with specific reference to the control of locomotion *Annu Rev Neurosci* 1 61, 1978

Stein, R B., and Lee, R. G Tremor and clonus *In* Brooks, V. B (ed.), Handbook of Physiology Sec 1, Vol II. Bethesda, American Physiological Society, 1981, p 325

Talbott, R. E., and Humphrey, D R. (eds.) Posture and Movement New York, Raven Press, 1979

Valentinuzzi, M The Organs of Equilibrium and Orientation as a Control System New York, Harwood Academic Publishers, 1980

Wilson, V., and Jones, G M Mammalian Vestibular Physiology New York, Plenum Press, 1979

Wilson, V J., and Peterson, B W Peripheral and central substrates of vestibulospinal reflexes *Physiol Rev*, *58* 80, 1978

Wilson, V J., and Peterson, B W Vestibulospinal and reticulospinal systems *In* Brooks, V B (ed.) Handbook of Physiology Sec 1, Vol. II Bethesda, American Physiological Society, 1981, p 667.

Young, L. R Cross coupling between effects of linear and angular acceleration on vestibular nystagmus *Bibl. Ophthalmol*, *82* 116, 1972

Young, L R. Perception of the body in space Mechanisms *In* Darian-Smith, I (ed) Handbook of Physiology Sec. 1, Vol III. Bethesda, American Physiological Society, 1984 p 1023.

53

Cortical and Cerebellar Control of Motor Functions

In preceding chapters we have been concerned with many of the subconscious motor activities integrated in the spinal cord and brain stem, especially those responsible for posture and equilibrium. In the present chapter we will discuss the control of motor function by the cerebral cortex and cerebellum as well as their relationship to the basal ganglia and the other lower centers. Much of this control is "voluntary" in contradistinction to the subconscious control effected by the lower centers. Yet we will also see that at least some motor functions of the cerebral cortex and cerebellum are not entirely "voluntary."

THE MOTOR CORTEX—THE PRIMARY AND PREMOTOR AREAS

Figure 53–1 illustrates a broad area of the cerebral cortex that is concerned either with sensation from the somatic areas of the body or with control of body movement. The posterior part of this area, the *somatic sensory cortex*, we have already discussed. Lying directly anterior to the somatic sensory area in front of the central sulcus, as illustrated in the figure, and occupying approximately the posterior one half of the frontal lobes is the *motor cortex*. Nerve signals originating from this region cause muscle contractions in different parts of the body

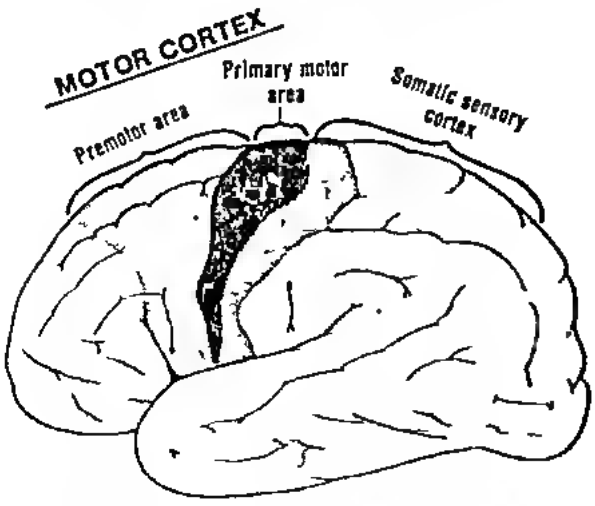


Figure 53–1. Relationship of the motor cortex to the somatic sensory cortex.

The motor cortex is divided into two separate divisions, the *primary motor area* and the *premotor area* The primary area contains very large *pyramidal motor neurons* that send their fibers all the way to the spinal cord through the *corticospinal tract* and therefore have almost direct communication with the anterior motor neucons of the cord for control of either individual muscles or small groups of muscles Even very weak electrical stimuli in this primary motor area usually will elicit a muscle contraction somewhere in the body. The primary motor cortex is also frequently referred to as *area IV* of the cortex because this area containing the large pyramidal cells is area IV in Brodmann's histological classification of the different cortical areas, as will be explained in Chapter 55.

The *premotor area* lies anterior to the primary motor area and is also illustrated in Figure 53–1. This area is frequently called areas VI and VIII because it occupies both these areas in the Brodmann classification of brain topology. The premotor cortex has very few neurons that project nerve fibers directly to the spinal cord. Instead, most of the nerve signals generated in this area cause more complex muscle movements, usually involving groups of muscles performing some specific task, rather than individual muscles. To achieve these results, the premotor area mainly sends its signals into the primary motor cortex to excite multiple groups of muscles. Some of these signals pass directly to the motor cortex through subcortical nerve fibers, but the premotor cortex also has extensive connections with the basal ganglia and cerebellum, both of which transmit signals back by way of the thalamus to the motor cortex. Thus

the premotor cortex, the basal ganglia, the cerebellum, and the primary motor cortex constitute a complex overall system for voluntary control of muscle activity.

THE TOPOGRAPHICAL MAP OF THE PRIMARY MOTOR AREA

The neurons in each small point area of the primary motor cortex project to a single muscle or small group of muscles in a specific part of the body. Figure 53–2 illustrates four separate so-called *topographical maps*, two sensory and two motor, that have been discovered in the brain of a monkey. These maps show specific points in the cerebral cortex connecting with the different areas of the body. The areas labeled *Sm I* and *Sm II* are the somatic sensory areas I and II that were described in Chapter 49. The other two topographical maps, *Ms I* and *Ms II*, located more anteriorly in the brain, are topographical maps for the fiber projections from the motor cortex outward to the different body areas. The dark area, Ms I, is the *primary motor cortex*, and the lighter area, Ms II, is called the *supplemental motor area*, which we shall discuss later.

Most of the primary cortex lies in the gyrus immediately anterior to the central sulcus, called the *precentral gyrus*, and, as illustrated in the figure, the head is represented most laterally in this gyrus whereas the tail end of the animal is represented most medially. In fact, the most caudal portions of the animal are represented in the portion of the precentral gyrus that curves downward into the longitudinal fissure and lies on the medial surface of the cerebral hemisphere, as illustrated in the folded-up upper portion of Figure 53–2 that shows this medial cerebral surface. Note in this figure that some parts of the body have much greater representation

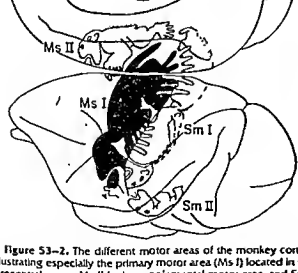


Figure 53–2. The different motor areas of the monkey cortex, illustrating especially the primary motor area (Ms I) located in the precentral gyrus. Ms II is the supplemental motor area, and Sm I and Sm II are the two somatic sensory areas. (The upper portion of the figure illustrates the medial surface of the cerebral hemisphere.) (from Woolsey, in Schaltenbrand and Woolsey (eds.) Cerebral Localization and Organization. University of Wisconsin Press, 1964.)

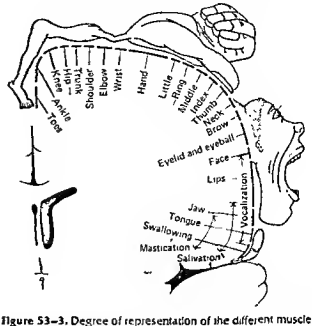


Figure 53–3. Degree of representation of the different muscles of the body in the motor cortex. (From Penfield and Rasmussen. The Cerebral Cortex of Man. A Clinical Study of Localization of Function. New York. Macmillan Company, 1968.)

in the motor cortex than other parts, especially the hands, the feet, and the digits.

The Motor Topographical Map in the Human Being. The topographical map of the motor cortex in the human being is quite different from that of lower animals. The reason for this is that the human being has developed two important capabilities involving the motor cortex that are not found in lower animals. These are (1) an exceptional capability to use the hand, the fingers, and the thumb to perform highly dexterous manual tasks, and (2) use of the mouth, lips, tongue, and facial muscles to talk. Therefore, there are very high degrees of representation of the hand, mouth, and facial regions in the human motor cortex.

Figure 53–3 illustrates the degrees of representation of the different muscle areas of the body in the motor cortex as mapped by Penfield and Rasmussen. This was done by stimulating the different areas of the motor cortex in human beings undergoing neurosurgical operations. Note that more than one half of the entire primary motor cortex is concerned with controlling the hands and the muscles of speech. Point stimulations in these areas of the motor cortex will cause contraction of a single muscle or even a portion of a single muscle. On the other hand, in those areas of the primary cortex with less intense degree of representation such as in the trunk area, electrical stimulation will usually contract a group of muscles instead.

The Supplemental Motor Area

Area Ms II in Figure 53–2, located on the medial surface of the frontal lobe slightly anterior to the primary motor cortex, is called the *supplemental motor area*. This area requires considerably stronger electrical stimuli to cause muscle contraction than does the primary motor area. Also, the movements involve coordinate contractions of many muscles in contradistinction to the much more discrete movements elicited from the primary area. Furthermore, many of the movements are

bilateral rather than unilateral, and the contractions occur most often in the trunk or proximal portions of the limbs, causing the animal to position itself in some special attitude. Also, there may be rotation of the head, movement of the eyes, vocalization, or yawning

COMPLEX MOVEMENTS ELICITED BY STIMULATING THE PREMOTOR AREA

Electrical stimulation of the premotor area will often elicit complex rontractions of groups of muscles. Occasionally, vocalization occurs, or rhythmic movements such as alternate thrusting of a leg forward and backward, coordinate moving of the eyes, chewing, swallowing, or contortions of parts of the body into different postural positions.

Some neurophysiologists have called this area the *motor association area* and have ascribed special capabilities to it to control coordinated movements involving many muscles simultaneously In fact, it is peculiarly organized to perform such a function for the following reasons: (1) it has long subcortical neuronal connections with the sensory association areas of the parietal lobe; (2) it has direct subcortical connections with the primary motor cortex; (3) it connects with areas in the thalamus contiguous with the thalamic areas that connect with the primary motor cortex; and (4) most important of all, it has abundant direct connections with the basal ganglia and the cerebellum, both of which feed back to the primary motor area through the thalamus

Broca's Area and Speech. Figure 53–4 illustrates a premotor area lying immediately anterior to the primary motor rortex and immediately above the sylvian fissure labeled "word formation" This region is called *Broca's area* Damage to it does not prevent a person from vocalizing, but it does make it impossible for the person to speak whole words other than simple utterances such as "no" or "yes." A closely associated rortical area also causes appropriate respiratory function : that respiratory activation of vocal rords can occ aimultaneously with the movements of the mou and tongue during speech. Thus, the premot activities that are related to Broca's area a highly complex.

The Voluntary Eye Movement Field. Imm diately above Broca's area is a locus for controllir eye movements. Damage to this area prevents person from voluntarily moving the eyes towa different objects. Instead, the eyes tend to lock c specific objects, an effect rontrolled by signals fro the occipital region, as explained in Chapter 6 This frontal area also rontrols eyelid movemen such as blinking.

Head Rotation Area. Still slightly higher i the motor association area, electrical stimulatic will elicit head rotation. This area is closely ass ciated with the eye movement field and is presun ably related to directing the head toward differe objects

Area for Hand Skills. In the frontal area in mediately anterior to the primary motor rortex f the hands and fingers is a region neurosurgeo have called an area for hand skills That is, whe tumors or other lesions cause destruction in th area, the hand movements become incoordina and nonpurposeful, a condition called *mot apraxia*

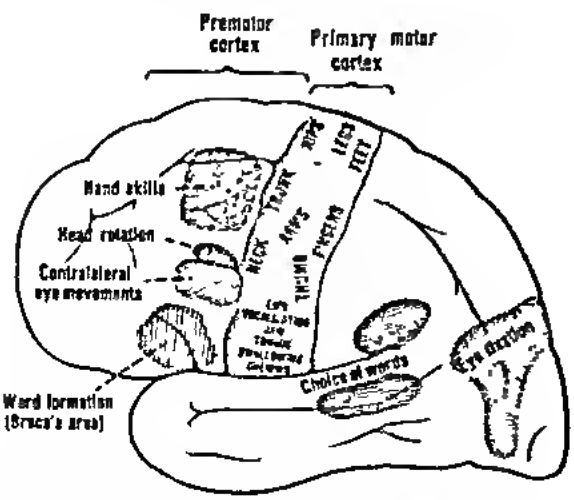


Figure 53–4. Representation of the different muscles of the body in the motor cortex and location of other cortical areas responsible for certain types of motor movements.

FIBER PATHWAYS TO AND FROM THE MOTOR CORTEX

The functions of the motor cortex are controll mainly by the somatic sensory system but also to lesser extent by the other sensory systems such hearing and vision. Once the sensory information derived from these sources, the motor cortex operates association with the basal ganglia and cerebellum process the information and to determine the appropr ate course of motor action. Therefore, there are mar two-way fiber pathways between the motor cortex the one hand and the sensory cortex, the basal gangl and the cerebellum on the other hand In addition, t motor cortex gives rise to an especially important mot output pathway directly to the spinal rord through t pyramidal tract, as well as many other shorter pathwa terminating primarily in the brain stem. Thus, t important pathways subserving all these functions a the following

Afferent Fibers That Excite the Motor Cortex

The motor rortex is stimulated to action by signa from many different sources, including:

1. Subcortical fibers from adjacent regions of t cortex, especially from the somatic sensory areas ar from the frontal areas—also subcortical fibers from t visual and auditory cortices
2. Subcortical fibers that pass through the corpt callosum from the opposite cerebral hemisphere. The fibers connect corresponding areas of the motor cortic in the two sides of the brain.

3. Somatic sensory fibers derived directly from the ventrobasal complex of the thalamus. These transmit mainly cutaneous tactile signals and joint and muscle signals.

4. Tracts from the ventrolateral and ventroanterior nuclei of the thalamus, which in turn receive tracts from the cerebellum and the basal ganglia These tracts provide signals that are necessary for coordination between the functions of the motor cortex, the basal ganglia, and the cerebellum.

5. Fibers from the nonspecific nuclei of the thalamus. These fibers probably control the general level of excitability of the motor cortex in the same manner that they also control the general level of excitability of most other regions of the cerebral cortex.

Efferent Fibers from the Motor Cortex

The Pyramidal Tract (Corticospinal Tract). The most important output pathway from the motor cortex is the *pyramidal tract*, also called the *corticospinal tract*, which is illustrated in Figure 53–5.

The pyramidal tract originates about 60 per cent from the primary motor cortex, 20 per cent from the premotor cortex, and 20 per cent from the somatic sensory areas posterior to the central sulcus. After leaving the cortex it passes through the posterior limb of the internal capsule (between the caudate nucleus and the putamen of the basal ganglia) and then downward through the brain stem, forming the *pyramids of the medulla*. By far the majority of the pyramidal fibers then cross to the opposite side and descend in the *lateral corticospinal tracts* of the cord, finally terminating principally on the interneurons in the intermediate regions of the cord gray matter However, some of the fibers in human beings (but not in most lower animals) terminate directly on the anterior motor neurons.

A few of the fibers do not cross to the opposite side in the medulla but pass ipsilaterally down the cord in the *ventral corticospinal tracts*, but these fibers also cross mainly to the opposite side of the cord either in the neck or the upper thoracic region

The most impressive fibers in the pyramidal tract are a population of large myelinated fibers with mean diameter of about 16 microns These originate from the *giant pyramidal cells*, also called *Betz cells*, that are found only in the primary motor cortex. These cells are about 60 microns in diameter, and their fibers transmit nerve impulses to the spinal cord at a velocity of about 70 meters per second, the most rapid rate of transmission of any signals from the brain to the cord. There are approximately 34,000 of these large fibers from the Betz cells in each corticospinal tract However, the total number of fibers in each corticospinal tract is more than a million, so these large fibers represent only 3 per cent of all of them. The other 97 per cent are mainly fibers smaller than 4 microns in diameter

Other Fiber Pathways from the Motor Cortex. The motor cortex gives rise to very large numbers of direct fibers from the cortex or collaterals from the pyramidal tract that go to the more basal regions of the cerebrum and also into the brain stem, including the following:

1. The axons from the giant Betz cells send short collaterals back to the cortex itself It is believed that these collaterals mainly inhibit adjacent regions of the cortex when the Betz cells discharge, thereby "sharpening" the boundaries of the excitatory signal

2 A large body of fibers passes into the *caudate nucleus* and *putamen*. From here additional pathways extend through several neurons into the brain stem, as discussed in the previous chapter.

3 A moderate number of fibers pass to the *red nuclei*. From these, additional fibers pass down the cord through the *rubrospinal tract*

4 A moderate number of fibers deviate into the *reticular substance* of the mesencephalon; from here signals go to the cord via *reticulospinal tracts* and others go to the cerebellum via *reticulocerebellar tracts*.

5. A tremendous number of fibers synapse in the pontile nuclei, which give rise to the *pontocerebellar fibers*. Thus, whenever signals are transmitted from the motor cortex through the pyramidal tract, simultaneous signals are transmitted into the cerebellar hemispheres.

6. Collaterals also terminate in the *inferior olivary nuclei*, and from here secondary *olivocerebellar fibers* transmit signals to many areas of the cerebellum.

Thus, the basal ganglia, the brain stem, and the cerebellum all receive strong signals from the pyromidal tract every time a signal is transmitted down the spinal cord to cause a motor activity.

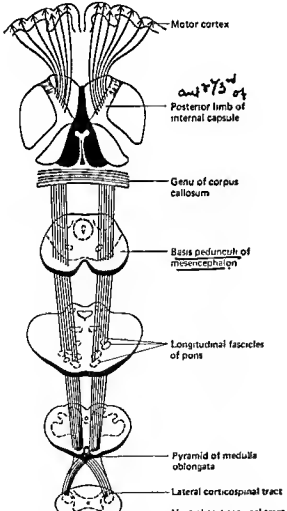


Figure 53–5. The pyramidal tract. (From Ranson and Clark: Anatomy of the Nervous System. Philadelphia, W. B. Saunders Company, 1959.)

THE RED NUCLEUS AND THE RUBROSPINAL TRACT—THEIR RELATIONSHIP TO THE PYRAMIDAL SYSTEM

The red nucleus, illustrated in Figure 53–6, is located in the mesencephalon and functions in close association with the pyramidal tract. This nucleus has two parts, a superior portion, called the *parvocellular portion*, that is composed of small neurons, and an inferior portion, called the *magnocellular portion*, that contains many large neurons. The large neurons of the magnocellular portion give rise to the *rubrospinal tract* that crosses to the opposite side in the lower brain stem and follows a course parallel to the corticospinal tract into the lateral columns of the spinal cord. This tract partially overlaps the corticospinal tract but on the average lies slightly anterior to it. The rubrospinal fibers terminate mainly on the interneurons of the intermediate areas of the cord gray matter along with the corticospinal fibers, and a few of the rubrospinal fibers also terminate on the anterior motor neurons, along with some of the corticospinal fibers.

The red nucleus receives two major input pathways. One of these is *from the motor cortex via the corticorubral tract*, terminating mainly in the magnocellular portion of the red nucleus and thus stimulating the fibers of the rubrospinal tract. The second source of input fibers is *from the cerebellum*, which provides fibers to both the parvocellular and magnocellular portions of the red nucleus.

Function of the Corticorubrospinal System. The magnocellular portion of the red nucleus has a somatotopical representation of all the muscles of the body, as is true of the motor cortex. Therefore, stimulation of a single point in this portion of the red nucleus will cause contraction of either a single muscle or small group of muscles. However, the fineness of representation of the different muscles is far less developed than is true in the motor cortex. This is especially true in human beings who have a relatively small red nucleus.

The corticorubrospinal pathway serves as an accessory route for the transmission of relatively discrete signals from the motor cortex to the spinal cord. When the pyramidal fibers are destroyed without destroying this other pathway, discrete movements can still occur, except that the movements of the fingers and hands are considerably impaired. Wrist movements are still well developed, which is not true when the corticorubrospinal pathway is also blocked. Therefore, the pathway through the red nucleus to the spinal cord is associated far more with the pyramidal system than with the vestibuloreticulospinal system that controls mainly the axial and girdle muscles of the body. Furthermore, the rubrospinal tract lies in the lateral columns of the spinal cord, along with the corticospinal tracts, and terminates more on the interneurons and motor neurons that control the distal muscles of the limbs. Therefore, the corticospinal and rubrospinal tracts together are frequently called the *lateral motor system of the cord*, in contradistinction to the vestibuloreticulospinal system that lies mainly medially in the cord and is called the *medial motor system of the cord*.

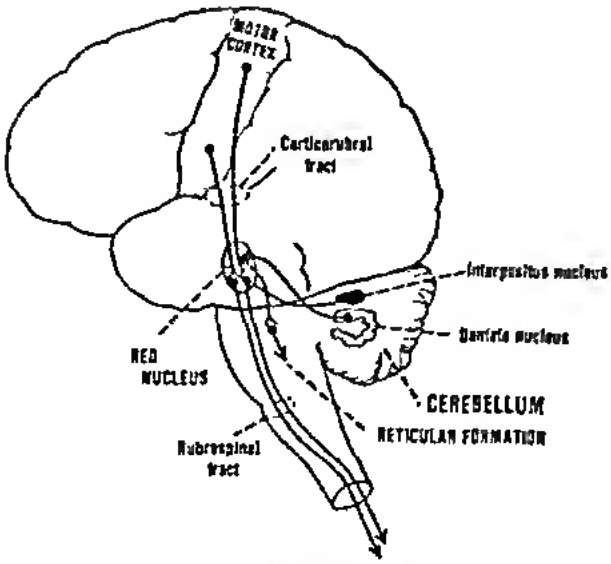


Figure 53–6. The corticorubrospinal pathway for motor control, showing also the relationship of this pathway to the cerebellum.

THE EXTRAPYRAMIDAL SYSTEM

The term *extrapyramidal motor system* is widely used in clinical circles to denote all those portions of the brain and brain stem that contribute to motor control but that are not part of the direct pyramidal system. This includes the basal ganglia, the reticular formation of the brain stem, the vestibular nuclei, and often the red nuclei as well. However, this is such an all-inclusive and diverse group of motor control areas that it is difficult to ascribe specific neurophysiological functions to the extrapyramidal system as a whole. For this reason, we have discussed the functions of the separate portions of the extrapyramidal system individually.

EXCITATION OF THE SPINAL CORD BY THE PRIMARY MOTOR CORTEX AND THE RED NUCLEUS

Vertical Columnar Arrangement of the Neurons in the Motor Cortex. In Chapters 49 and 60 it is pointed out that the cells in the somatic sensory cortex and visual cortex—and perhaps in all other parts of the brain as well—are organized in vertical columns of cells. In a like manner, the cells of the motor cortex are also organized in vertical columns—columns 0.3 to 1 mm in diameter and having perhaps 50,000 to 150,000 neurons in each column.

Each column of cells functions as a unit, stimulating either a single muscle or a group of synergistic muscles. Also, each column of cells is arranged in six distinct layers, like the arrangement throughout almost all the cerebral cortex. The pyramidal cells that give rise to the corticospinal fibers all lie in the fifth layer of cells from the

cortical surface, whereas the input signals to the column of cells all enter layers 2 through 4. The sixth layer gives rise mainly to fibers that communicate with other regions of the cerebral cortex itself.

Function of Each Column of Neurons. The neurons of each column operate as an integrative processing system, utilizing information from multiple input sources to determine the output response from the column. In addition, each column can function as an amplifying system so that large numbers of pyramidal fibers to the same muscle or to synergistic muscles can be stimulated simultaneously. This is important because stimulation of a single pyramidal cell can rarely excite a muscle. Instead, as many as perhaps 50 to 100 pyramidal cells need to be excited simultaneously or in rapid succession to achieve muscle contraction.

Dynamic and Static Signals Transmitted by the Pyramidal Neurons. Characteristically, when a weak signal is transmitted to a muscle the muscle develops its force of contraction very slowly. However, if a strong signal is sent at first to cause initial rapid contraction, then a much weaker signal can maintain the contraction for long periods thereafter. This is the manner in which excitation for causing muscle contractions is usually provided. To do this, each column of cells excites two separate populations of pyramidal cell neurons, one called the *dynamic neurons* and the other the *static neurons*. The dynamic neurons are excessively excited for a short period of time at the beginning of the contraction, causing the initial development of force. Then the static neurons fire at a much slower rate, but they continue at this slow rate indefinitely to maintain the force of contraction as long as the contraction is required.

The neurons of the red nucleus have similar dynamic and static characteristics, except that more dynamic neurons are in the red nucleus and more static neurons in the primary motor cortex. This perhaps relates to the fact that the red nucleus is closely allied to the cerebellum, and we shall learn later in this chapter that the cerebellum also plays an important role in the rapid initiation of muscle contraction.

Somatic Sensory Feedback to the Motor Cortex. When nerve signals from the motor cortex cause a muscle to contract, somatic sensory signals return from the activated region of the body to the neurons in the motor cortex that are causing the action. Most of these somatic sensory signals arise in the muscle spindles or in the tactile receptors of the skin overlying the muscles. In general, the somatic signals cause a positive feedback enhancement of the muscle contraction in the following ways: In the case of the muscle spindles, if the fusimotor muscle fibers in the spindles contract more than the large skeletal muscle itself does, then the spindles become excited and the signals from these spindles stimulate the pyramidal cells in the motor cortex, which further excites the muscle, helping its contraction catch up with the contraction of the spindles. In the case of the tactile receptors, if the muscle contraction causes compression of the skin against an object, such as compression of the fingers around an object that is being grasped, the signals from these receptors cause further excitement of the muscles and therefore increase the muscle contraction—such as increasing the tightness of the grasp of the hand.

Stimulation of the Spinal Motor Neurons

Figure 53–7 shows a segment of the spinal cord, illustrating multiple motor tracts entering the cord from the brain and also showing a representative anterior motor neuron. The corticospinal tract and the rubrospinal tract lie in the dorsal portions of the lateral columns. Most of their fibers terminate on interneurons in the intermediate area of the cord gray matter; this is especially so in the trunk, leg, and arm areas of the cord. However, in the cervical enlargement of the cord where the hands and fingers are represented, the motor neurons supplying the hands and fingers lie almost entirely in the lateral portions of the anterior horns. And large numbers of both corticospinal and rubrospinal fibers terminate directly on these anterior motor neurons, thus allowing a direct route from the brain for activating muscle contraction. This is in keeping with the fact that the primary motor cortex has an extremely high degree of representation for fine control of hand, finger, and thumb actions.

Patterns of Movement Elicited by Spinal Cord Centers. From Chapter 51, recall that the spinal cord can provide specific reflex patterns of movement in response to sensory nerve stimulation. Many of these patterns are also important

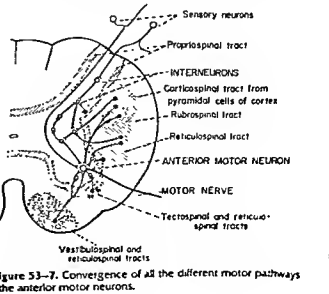


Figure 53–7. Convergence of all the different motor pathways on the anterior motor neurons.

when the anterior motor neurons are excited by signals from the brain. For instance, the stretch reflex is functional at all times, helping damp the motor movements initiated from the brain and probably providing at least part of the motive power required to cause the muscle contractions employing the servo-assist mechanism that was described in Chapter 51.

Also, when a brain signal excites an agonist muscle, it is not necessary to transmit an inverse signal to the antagonist at the same time; this transmission will be achieved by the reciprocal innervation circuit that is always present in the cord for coordinating the functions of antagonistic pairs of muscles

Finally, parts of the other reflex mechanisms, such as withdrawal, stepping and walking, scratching, postural mechanisms, and so forth, can be activated by "command" signals from the brain. Thus, very simple signals from the brain can lead, at least theoretically, to many of our normal motor activities, particularly for such functions as walking and the attainment of different postural attitudes of the body

Effect of Lesions in the Motor Cortex—The "Stroke"

The motor cortex is frequently damaged, especially by the common abnormality called a "stroke," which is caused either by a ruptured blood vessel that allows hemorrhage into the brain or by thrombosis of one of tha major arteries supplying the brain, in either case causing loss of blood supply to the cortex. Also, experiments have been performed in animals to remove selectively different parts of the motor cortex

Removal of the Primary Motor Cortex (the Area Pyramidalis). Removal of a very small portion of the primary motor cortex—the area that contains the giant Betz pyramidal cells—in a monkey causes varying degrees of paralysis of the represented muscles. If the sublying caudate nucleus and the adjacent premotor area are not damaged, gross postural and limb "fixation" movements can still be performed, but the animal *loses voluntary control of discrete movements of the distal segments of the limbs—especially of the hands and fingers* This does not mean that the muscles themselves cannot contract but that the animal's ability to control the fine movements is gone.

From these results one can conclude that the area pyramidalis is essential for voluntary initiation of finely controlled movements, especially of the hands and fingers

Muscle Spasticity Caused by Large Lesions of the Motor Cortex and Basal Ganglia. Ablation of the primary motor cortex alone or interruption of the pyramidal tract causes hypotonia, not spasticity, because the primary motor cortex normally exerts a continual tonic stimulatory effect on the motor neucons of the spinal cord, when they are removed, hypotonia results.

On the other hand, most lesions of the motor cortex, especially those caused by a stroke, involve both the motor cortex itself and deeper structures of the cerebrum as well, especially the basal ganglia. In these instances, muscle spasm almost invariably occurs in the afflicted muscle areas on the opposite side of the body (because all the motor pathways cross to the opposite side). Obviously, this spasm is not caused by loss of either the primary motor cortex or blockage of the pyramidal fibers to the cord. Instead, it is believed to be caused mainly by the damage in the basal ganglia In fact, the greater the damage to the basal ganglia, the greater the degree of spasm.

It will be recalled that a major share of the output signals from the basal ganglial system are inhibitory. Much of this inhibition feeds back to the primary motor cortex to inhibit the pyramidal system itself, but still additional inhibitory signals pass into the brain stem to inhibit the reticular formation and vestibular nuclei. Therefore, release of these other motor systems fcom the basal ganglial inhibition allows excess excitatory motor activity, leading to the muscle spasm

The muscle spasm that results fcom large lesions of the motor cortex and basal ganglia is different fcom the spasticity that occurs in decerebrate rigidity. In decerebrate rigidity the spasm occurs almost exclusively in the antigravity muscles, activated to a great extent by excess activity in the muscle spindle stretch reflexes; it is called *extensor spasm* On the other hand, the spasm resulting from large cortical lesions usually involves the flexor muscles as well and is activated to a major extent by the alpha motor neurons to the muscles, in addition to activation of the muscle spindle system This causes intense stiffening of the limbs and other parts of the body

The Babinski Sign. Destruction of the foot region of the area pyramidalis or transection of tha foot portion of the pyramidal tract causes a peculiar response of the foot called the *Babinski sign*. This response is demonstrated when a firm tactile stimulus is applied to the lateral sole of the foot The great toe extends upward and the other toes fan outward. This is in contradistinction to the normal effect in which all the toes bend downward The Babinski sign does not occur when damage occurs in the extrapyramidal portions of the motor contcol system without involving the pyramidal tract. Therefore, the sign is used clincially to detect damage specifically in the pyramidal portion of the motor control system.

The cause of the Babinski sign is believed to be the following The pyramidal tract is a major controller of muscle activity for performance of voluntary, purposeful activity. On the other hand, the extrapyramidal system, a much older motor control system, is concerned to a great extent with protection. Therefore, when only the extrapyramidal system is functional, stimuli to the bottom of the feet cause a typical withdrawal pcotective type of reflex which is expressed by the upturned great toe and fanning of the other toes. But, when the pyramidal system is also fully functional, it suppresses the pcotective reflex and instead excites a higher order of motor function, including the normal effect of causing downward bending of the toes in response to sensory stimuli from the bottom of the feet.

THE CEREBELLUM AND ITS MOTOR FUNCTIONS

The cerebellum has long been called a *silent area* of the brain principally because electrical excitation of this structure does not cause any sensation

and rarely any motor movement. However, as we shall see, removal of the cerebellum does cause the motor movements to become highly abnormal. The cerebellum is especially vital to the control of very rapid muscular activities such as running, typing, playing the piano, and even talking. Loss of this area of the brain can cause almost total incoordination of these activities even though its loss causes paralysis of no muscles.

But how is it that the cerebellum can be so important when it has no direct capability of causing muscle contraction? The answer to this is that it both helps *plan the motor activities* and also *monitors and makes corrective adjustments in the motor activities elicited by other parts of the brain.* It receives continuously updated information on the desired program of muscle contractions from the motor control areas of the other parts of the brain. And it receives continuous information from the peripheral parts of the body to determine the instantaneous status of each part of the body—its position, its rate of movement, forces acting on it, and so forth. It is believed that the cerebellum *compares* tha actual instantaneoua status of each part of the body as depicted by the peripheral information with tha statua that is intended by tha motor system. If the two do not compare favorably, then appropriate corrective signala are transmitted instantaneously back into the motor system to increase or decrease the levels of activation of tha specific muscles.

Since tha cerebellum must make major motor correctiona extremely rapidly *during the course of motor movements,* a very extensive and rapidly acting cerebellar input system is required both from the peripheral parts of the body and from the cerebral motor areas. Also, an extensive output system feeding equally as rapidly into the motor system is necessary to provide the necessary corrections of the motor signals.

THE ANATOMICAL FUNCTIONAL AREAS OF THE CEREBELLUM

Anatomically, the cerebellum is divided into three separate lobes by two deep fissures, as shown in Figure 53–8: (a) the *anterior lobe,* (b) the *posterior lobe,* and (c) the *flocculonodular lobe.* The flocculonodular lobe is the oldest of all portions of the cerebellum; it developed along with (and functions with) the vestibular system in controlling equilibrium, as was discussed in the previous chapter. Because of its ancient heritage, it is frequently called the *archicerebellum.* The anterior lobe and part of the midportion of the posterior lobe are also old; this is called the *paleocerebellum.* On the other hand, almost 90 per cent of the posterior lobe is recent in origin and is especially highly developed in primates and human beings; this is called the *neocerebellum.*

The Longitudinal Functional Divisions of the Anterior and Posterior Lobes. From a functional point of view, the anterior and posterior lobes are organized not by lobes but instead along the longitudinal axis, as illustrated in Figure 53–9 which shows the human

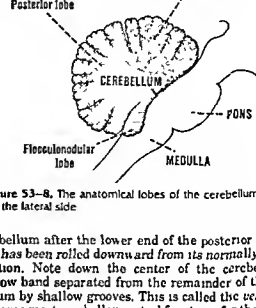


Figure 53–8. The anatomical lobes of the cerebellum as seen from the lateral side

cerebellum after the lower end of the posterior cerebellum has been rolled downward from its normally hidden position. Note down the center of the cerebellum a narrow band separated from the remainder of the cerebellum by shallow grooves. This is called the *vermis.* In this area most cerebellar control functions for the muscle movements of the axial body, the neck, and the shoulders and hips are located.

To each side of the vermis is a large, laterally protruding *cerebellar hemisphere,* and each of these hemispheres is divided into an *intermediate zone* and a *lateral zone.* The intermediate zone of the hemisphere is concerned with the control of muscular contractions in the distal portions of both the upper and lower limbs, especially of the hands and fingers and feet and toes. On the other hand, the lateral zone of the hemisphere operates at a much more remote level, for this area seems to join into the overall planning of sequential motor movements. Without this lateral zone, most discrete motor activities of the body lose their appropriate timing and therefore become highly incoordinate, as we shall discuss more fully later.

Topographical Representation of the Body in the Cerebellum. In the same manner that the sensory cortex, the motor cortex, the basal ganglia, the red nuclei, and the reticular formation all have topograph-

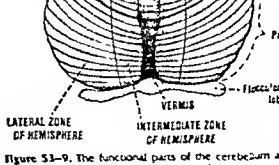


Figure 53–9. The functional parts of the cerebellum as seen from the posteroinferior view, with the inferomost portion of the cerebellum rolled outward to flatten the surface

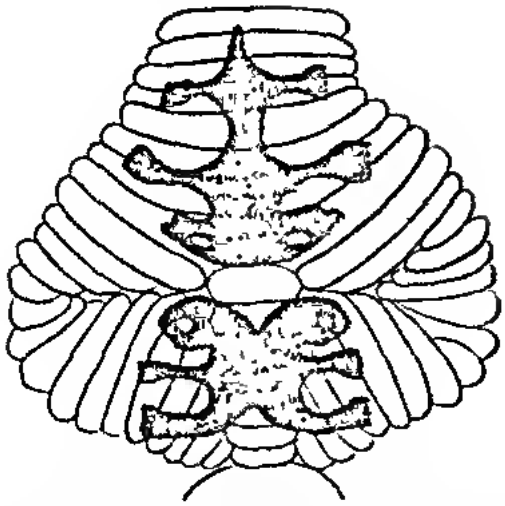


Figure 53–10. The sensory projection areas, called homunculi, on the cortex of the cerebellum (From Snider *Sci Amer* 199:4 1958 © 1958 by Scientific American Inc. All rights reserved.)

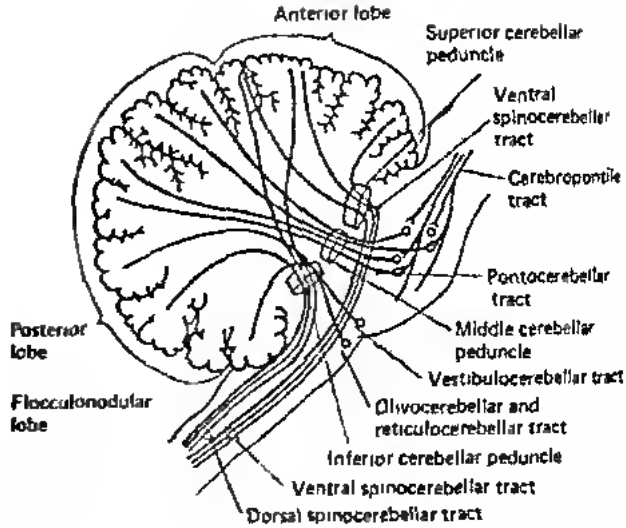


Figure 53–11. The principal afferent tracts to the cerebellum.

ical representations of the different parts of the body, so also is this true for parts of the cerebellum. Figure 53–10 illustrates two separate such representations in a small monkey, showing one to be located in the anterior lobe and the other in the posterior lobe. Note that the axial portions of the body lie in the vermal part of the cerebellum whereas the limbs and facial regions lie in the intermediate zones of the two hemispheres. These topographical representations receive afferent nerve fibers from all the respective parts of the body. In turn, they send motor signals into the same respective topographical areas of the motor cortex, the basal ganglia, the red nucleus, and the reticular formation.

However, note that the large lateral portions of the cerebellar hemispheres *do not* have topographical representations of the body. These areas of the cerebellum connect mainly with the association areas of the brain, especially the premotor area of the frontal cortex and the somatic sensory and sensory association areas of the parietal cortex. Presumably this connectivity with the association areas allows the lateral portions of the cerebellar hemispheres to play important roles in planning and coordinating the sequential patterns of muscular activities.

The Input Pathways to the Cerebellum

Afferent Pathways from the Brain. The basic input pathways to the cerebellum are illustrated in Figure 53–11. An extensive and important afferent pathway is the *corticopontocerebellar pathway*, which originates mainly in the *motor cortex* but to a lesser extent in the sensory cortex as well and then passes by way of the *pontile nuclei* and *pontocerebellar tracts* to the contralateral hemisphere of the cerebellum. In addition, important afferent tracts originate in the brain stem; they include (a) an extensive *olivocerebellar* tract, which passes from the *inferior olive* to all parts of the cerebellum; this tract is excited by fibers from the *motor cortex*, the *basal ganglia*, widespread areas of the *reticular formation*, and the *spinal cord*, (b) *vestibulocerebellar fibers*, some of which originate in the vestibular apparatus itself and others from the vestibular nuclei; most of these terminate in the *flocculonodular lobe* and *fastigial nucleus* of the cerebellum; and (c) *reticulocerebellar fibers*, which originate in different portions of the reticular formation and terminate mainly in the midline cerebellar areas (the vermis).

Afferent Pathways from the Periphery. The cerebellum also receives important sensory signals directly from the peripheral parts of the body through four separate tracts, two of which are located dorsally in the cord and two ventrally. The two most important of these tracts are illustrated in Figure 53–12: the *dorsal spinocerebellar tract* and the *ventral spinocerebellar tract*. These two tracts originate in the sacral, lumbar, and thoracic segments of the cord. Similar tracts, not shown in Figure 53–12, originate in the neck segments of the cord and course roughly along with the dorsal and ventral spinocerebellar tracts in their passage to the

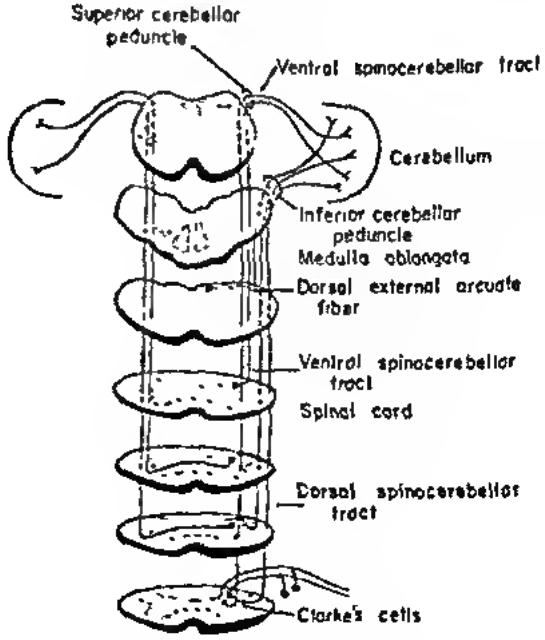


Figure 53–12. The spinocerebellar tracts.

cerebellum. These are the *cuneocerebellar tract* that joins the dorsal spinocerebellar tract and the *rostral spinocerebellar tract* that joins the ventral spinocerebellar tract. The dorsal tracts enter the cerebellum through the inferior cerebellar peduncle and terminate in the cerebellum on the same side as their origin. The two ventral tracts enter the cerebellum through the superior cerebellar peduncle, but they terminate in both sides of the cerebellum.

The signals transmitted in the dorsal spinocerebellar tracts come mainly from the muscle spindles and to a lesser extent from other somatic receptors throughout the body, such as from the Golgi tendon organs, the large tactile receptors of the skin, and the joint receptors. All these signals apprise the cerebellum of the momentary status of muscle contraction, degree of tension on the muscle tendons, positions and rates of movement of the parts of the body, and forces acting on the surfaces of the body.

On the other hand, the ventral spinocerebellar tracts receive less information from the peripheral receptors Instead, they are excited mainly by the motor signals arriving in the spinal cord from the brain through the corticospinal and rubrospinal tracts. Thus, this ventral fiber pathway tells the cerebellum that the motor signals have indeed arrived at the cord, and it also apprises the cerebellum of the intensity of the signals.

The spinocerebellar pathways can transmit impulses at velocities as great as 100 meters per second, which is the most rapid conduction of any pathway in the entire central nervous system. This extremely rapid conduction is important for the instantaneous apprisal of the cerebellum of changes that take place in the status of the body.

In addition to the signals in the spinocerebellar tracts, other signals are transmitted through the dorsal and dorsolateral columns to the medulla and then relayed from there to the cerebellum Likewise, signals are transmitted through the *spinoreticular pathway* to the reticular formation of the brain stem and through the *spino-olivary pathway* to the inferior olivary nucleus and then relayed from both these areas to the cerebellum. Thus, the cerebellum continually collects information about all parts of the body even though it is operating at a subconscious level.

Output Signals from the Cerebellum

The Deep Cerebellar Nuclei and the Efferent Pathways. Located deep in the cerebellar mass are three *deep cerebellar nuclei*—the *dentate, interpositus,* and *fastigial nuclei.* The *vestibular nuclei* in the medulla also function in some respects as if they were deep cerebellar nuclei because of their direct connections with the cortex of the flocculonodular lobe. All the deep cerebellar nuclei receive signals from two different sources: (1) the cerebellar cortex and (2) the sensory afferent tracts to the cerebellum. Each time an input signal arrives in the cerebellum, it divides and goes in two directions: (1) directly to one of the deep nuclei and (2) to a corresponding area of the cerebellar cortex overlying the deep nucleus; then, a short time later, the cerebellar cortex relays its output signals also to the same deep nucleus. Thus, all the input signals that enter the cerebellum eventually end in the deep nuclei. We shall discuss this circuit in greater detail later.

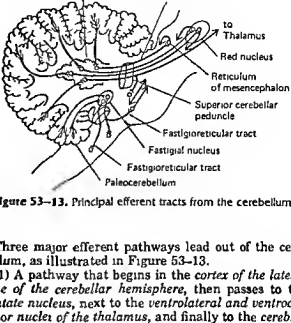


Figure 53–13. Principal efferent tracts from the cerebellum.

Three major efferent pathways lead out of the cerebellum, as illustrated in Figure 53–13.

(1) A pathway that begins in the *cortex of the lateral zone of the cerebellar hemisphere,* then passes to the *dentate nucleus,* next to the *ventrolateral and ventroanterior nuclei of the thalamus,* and finally to the *cerebral cortex* This pathway plays an important role in helping coordinate "voluntary" motor activities initiated by the cerebral cortex

(2) A pathway that originates in the *midline structures of the cerebellum* (the *vermis*) and then passes through the *fastigial nuclei* into the *medullary* and *pontile regions of the brain stem* This circuit functions in close association with the equilibrium apparatus to help control equilibrium and also, in association with the reticular formation of the brain stem, helps control the postural attitudes of the body It was discussed in detail in the previous chapter in relation to equilibrium.

(3) A pathway that originates in the *intermediate zone of the cerebellar hemisphere,* between the vermis and the lateral zone of the cerebellar hemisphere, then passes (a) through the *nucleus interpositus* to the *ventrolateral and ventroanterior nuclei of the thalamus,* and thence to the *cerebral cortex,* (b) to several *midline structures* of the *thalamus* and thence to the *basal ganglia,* and (c) to the *red nucleus* and *reticular formation* of the upper portion of the brain stem. This circuit is believed to coordinate mainly the reciprocal contractions of agonist and antagonist muscles in the peripheral portions of the limbs—especially in the hands, fingers, and thumbs.

THE NEURONAL CIRCUIT OF THE CEREBELLUM

The human cerebellar cortex is actually a large folded sheet, approximately 17 cm wide by 120 cm long, with the folds lying crosswise, as illustrated in Figures 53–9 and 53–10. Each fold is called a *folium.* And lying deep in the folded mass of cortex are the deep nuclei.

The Functional Unit of the Cerebellar Cortex—the Purkinje Cell. The cerebellum has approximately 30 million nearly identical functional units, one of which is illustrated to the left in Figure 53–14, shown mainly in red color. This

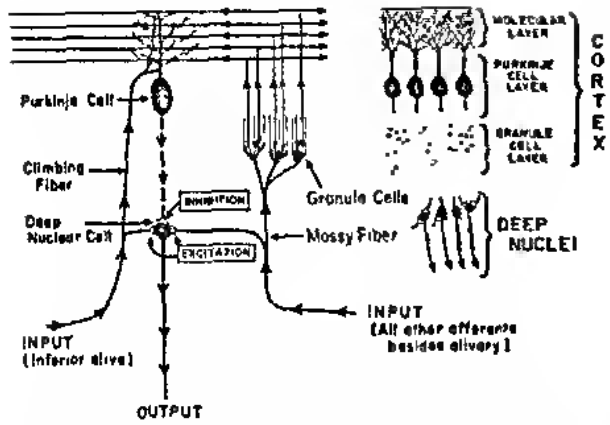


Figure 53–14. Basic neuronal circuit of the cerebellum showing excitatory pathways in color. At right are the three major layers of the cerebellar cortex and also the deep nuclei.

functional unit centers on the *deep nuclear cell* and on the *Purkinje cell*, of which there are also 30 million in the cerebellar cortex.

Note to the right in Figure 53–14 the three major layers of the cerebellar cortex: the *molecular layer*, the *Purkinje cell layer*, and the *granular cell layer*. Then, beneath these layers, the *deep nuclei* are located far within the center of the cerebellar mass.

The Neuronal Circuit of the Functional Unit. As illustrated in Figure 53–14, the output from the functional unit is from a deep nuclear cell. However, this cell is continually under the influence of both excitatory and inhibitory influences. The excitatory influences arise from direct connections with the afferent fibers that enter the cerebellum. The inhibitory influences arise entirely from the Purkinje cells in the cortex of the cerebellum.

The afferent inputs to the cerebellum are mainly of two types, one called the *climbing fiber type* and the other called the *mossy fiber type*. There is one climbing fiber for about 10 Purkinje cells. After sending collaterals to several deep nuclear cells, the climbing fiber projects all the way to the molecular layer of the cerebellar cortex where it makes about 300 synapses with the soma and dendrites of each Purkinje cell. This climbing fiber is distinguished by the fact that a single impulse in it will always cause a single, very prolonged, and peculiar oscillatory type of action potential in each Purkinje cell with which it connects. Another distinguishing feature of the climbing fibers is that *they all originate in the inferior olive of the medulla*, whereas the cerebellar afferent fibers from all other sources are almost entirely of the mossy type.

The mossy fibers also send collaterals to excite deep nuclear cells. Then these fibers proceed to the granular layer of the cortex where they synapse with hundreds of *granule cells*. These in turn send very small axons, less than 1 micron in diameter, up to the outer surface of the cerebellar cortex to enter the molecular layer. Here the axons divide into two branches that extend 1 to 2 millimeters in each direction parallel to the folia. There are literally millions of these *parallel nerve fibers* in each small segment of the cerebellar cortex (there are about 1000 granule cells for every Purkinje cell). It is into this molecular layer that the dendrites of the Purkinje cells project, and 80,000 to 200,000 of these parallel fibers synapse with each Purkinje cell; as these fibers pass along their 1 to 2 mm course, each of them contacts about 50 Purkinje cells. Yet, the mossy fiber input to the Purkinje cell is quite different from the climbing fiber input because stimulation of a single mossy fiber will never elicit an action potential in the Purkinje cell; instead, large numbers of mossy fibers must be stimulated simultaneously to activate the Purkinje cell. Furthermore, this activation usually takes the form of prolonged facilitation or excitation that, when it reaches threshold for stimulation, causes repetitive Purkinje cell firing of normal, short-duration action potentials rather than the single prolonged action potential occurring in response to the climbing fiber input.

Thus, the Purkinje cells are stimulated by two types of input circuits—one that causes a highly specific output in response to the incoming signal and the other that causes a less specific but tonic type of response. It should be noted that by far the greater proportion of the afferant input to the cerebellum is of the mossy fiber type, because this represents the afferent input from almost all the cerebellar afferent tracts besides those from the inferior olive.

Balance Between Excitation and Inhibition in the Deep Cerebellar Nuclei. The output signals from the Purkinje cells to the deep nuclei are entirely inhibitory. Therefore, referring again to the circuit of Figure 53–14, *one should note that direct stimulation of the deep nuclear cells by both the climbing and the mossy fibers excites them whereas the signals arriving from the Purkinje cells inhibit them.* Normally, there is a continual balance between these two effects so that the degree of output from the deep nuclear cell remains relatively constant at a moderate level of continuous stimulation. On the other hand, in the execution of rapid motor movements, the *timing* of the two effects on the deep nuclei is such that the excitation appears before the inhibition. Then a few milliseconds later inhibition occurs. In this way, there is first a very rapid excitotory signal fed back into the motor pathway to modify the motor movement, but this is followed within a few milliseconds by an inhibitory signal. This inhibitory signal resembles a "delay-line" negative feedback signal of the type that is very effective in providing *damping*. That is, when the motor system is excited, a negative feedback signal presumably occurs after a short delay to stop the muscle movement from overshooting its mark, which is the usual cause of oscillation.

Other Inhibitory Cells in the Cerebellar Cortex. In addition to the granule cells and Purkinje cells, three other types of neurons are also located in the cerebellar cortex: *basket cells, stellate cells,* and *Golgi cells.* All these are inhibitory cells with very short axons. Both the basket cells and the stellate cells are located in the molecular layer of the cortex, lying among and stimulated by the parallel fibers. These cells in turn send their axons at right angles across the parallel fibers and cause *lateral inhibition* of the adjacent Purkinje cells, thus sharpening the signal in the same manner that lateral inhibition sharpens the contrast of signals in many other areas of the nervous system. The Golgi cells lie in the Purkinje cell layer of the cortex, and their dendrites are also stimulated by the parallel fibers of the molecular layer, but the axons from these cells feed back to and inhibit the granule cells instead of the Purkinje cells. The function of this feedback is to limit the duration of the signal transmitted into the cerebellar cortex from the granule cells. That is, within a short fraction of a second after the granule cells are stimulated, they are then inhibited by the feedback. Therefore, a short, transient, pulselike signal, *not* a prolonged signal, is transmitted into the parallel fibers and thence to the Purkinje cells.

Special Features of the Cerebellar Neuronal Circuit. A special feature of the cerebellum is that there are no reverberatory pathways in the cerebellar neuronal circuits, so that the input-output signals of the cerebellum are very rapid transients that never persist for long periods of time.

Another special feature is that many of the cells of the cerebellum are constantly active. This is especially true of the deep nuclear cells; they continually send output signals to the other areas of the motor system. The importance of this is that *decrease* of the nuclear cell firing rate can provide an *inhibitory output signal* from the cerebellum, while an *increase* in firing rate can provide an *excitatory output signal.*

FUNCTION OF THE CEREBELLUM IN CONTROLLING MOVEMENTS

The cerebellum functions in motor control only in association with motor activities initiated elsewhere in the nervous system. These activities may originate in the spinal cord, in the reticular formation, in the basal ganglia, or in areas of the cerebral cortex. We will discuss, first, the operation of the cerebellum in association with the spinal cord and lower brain stem for control of postural movements and equilibrium and then discuss its function in association with the motor cortex for control of voluntary movements.

FUNCTION OF THE CEREBELLUM WITH THE SPINAL CORD AND BRAIN STEM TO CONTROL POSTURAL AND EQUILIBRIUM MOVEMENTS

The cerebellum originated phylogenetically at about the same time that the vestibular apparatus developed. Furthermore, as was discussed in the previous chapter, loss of the flocculonodular lobes of the cerebellum causes extreme disturbance of equilibrium. Yet, we still must ask the question, what role does the cerebellum play in equilibrium that cannot be provided by the other neuronal machinery of the brain stem? A clue is the fact that in persons with cerebellar dysfunction equilibrium is far more disturbed during performance of rapid motions than during stasis. This suggests that the cerebellum is especially important in controlling the balance between agonist and antagonist muscle contractions during rapid changes in body positions as dictated by the vestibular apparatuses. One of the major problems in controlling this balance is the time required to transmit position signals and kinesthetic signals from the different parts of the body to the brain. Even when utilizing the most rapidly conducting sensory pathways at 100 meters per second, as used by the spinocerebellar system, the delay for transmission from the feet to the brain is still 15 to 20 milliseconds. The feet of a person running rapidly can move as much as 10 inches during this time. Therefore, it is impossible for the brain to know at any given instant during rapid motion the exact position of the different parts of the body.

On the other hand, with appropriate neuronal circuitry, it would be possible for the cerebellum or some other portion of the brain to know how rapidly and in what direction a part of the body was moving 15 to 20 milliseconds earlier and then to predict from this information where the parts of the body should be at the present time. And this seems to be one of the major functions of the cerebellum.

As we have already discussed in relation to the neuronal circuitry of the cerebellum, there are abundant sensory pathways from the somatic areas of the body, especially from the muscles, joints, and skin surface, that feed both into the brain stem and into the older areas of the cerebellum—into the flocculonodular lobes through the vestibular nuclei and into the vermis and intermediate areas of the cerebellum through the dorsal and ventral spinocerebellar tracts and reticulocerebellar tracts. Also, the vestibular apparatus is located within a few centimeters of the flocculonodular lobes, allowing no more than a millisecond or so delay in transmission of the vestibular information.

Therefore, during the control of equilibrium, it is presumed that the extremely rapidly conducted vestibular apparatus information is used in a typical feedback control circuit to provide almost instantaneous correction of postural motor signals as necessary for maintaining equilibrium even during extremely rapid motion, including rapidly changing directions of motion. The feedback signals from the peripheral areas of the body help in this process, but their help is presumably contin-

gent upon some function of the cerebellum to compute positions of the respective parts of the body at any given time, despite the long delay time from the periphery to the cerebellum.

Relationship of Cerebellar Function to the Spinal Cord Stretch Reflex

One major component of cerebellar control of posture and equilibrium is an extreme amount of information transmitted from the muscle spindles to the cerebellum through the dorsal spinocerebellar tracts In turn, signals are transmitted into the brain stem through the cerebellar fastigial nuclei to stimulate the gamma efferent fibers that innervate the muscle spindles themselves. Therefore, a cerebellar stretch reflex occurs that is similar to but more complex than the spinal cord stretch reflex. It utilizes signals that pass all the way to the cerebellum and back again to the muscles. In general, this reflex adds additional support to the cord stretch reflex, but its feedback time is considerably longer, thus prolonging the effect. Through this feedback pathway many of the postural adjustments of the body are believed to occur.

FUNCTION OF THE CEREBELLUM IN VOLUNTARY MUSCLE CONTROL

In addition to the feedback circuitry between the body periphery and the cerebellum, an almost entirely independent feedback circuitry exists between the motor cortex and the cerebellum This is illustrated in its simplest form in Figure 53–15 and in a much more complex form, involving the basal ganglia also in the control circuit, in Figure 53–16 Most of the signals of this circuit pass from

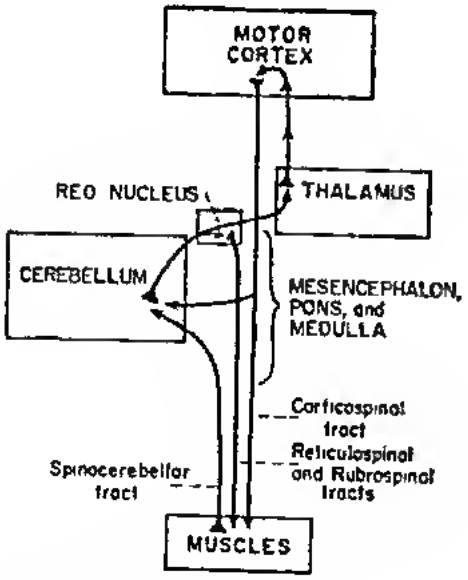


Figure 53–15. Pathways for cerebellar control of voluntary movements

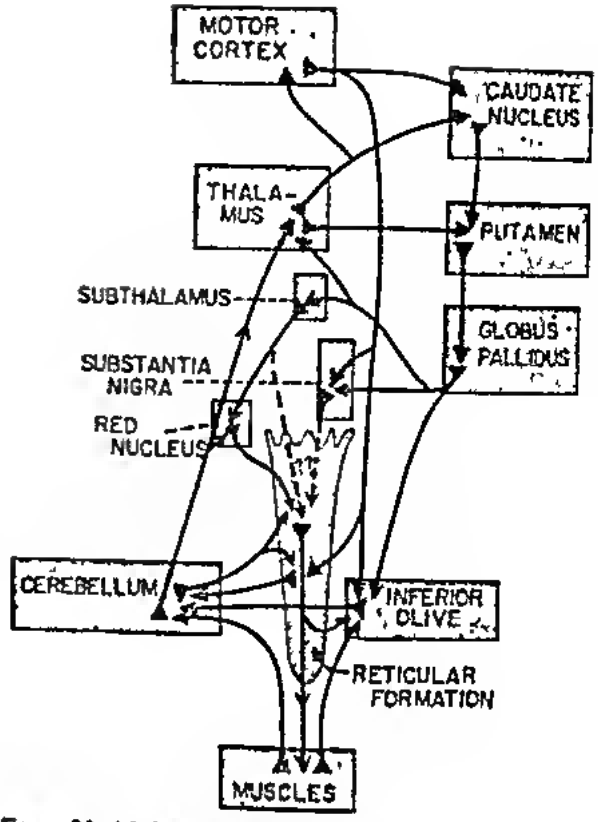


Figure 53–16. Pathways for cerebellar "error" control of involuntary movements

the motor cortex and adjacent cortical areas to the cerebellar hemispheres and then back to the cortex again, successively, through the dentata and interpositus cerebellar nuclei and the ventrolateral and ventroanterior nuclei of the thalamus. These circuits are not involved in the control of the axial and girdle muscles of the body. Instead, they serve two other motor control functions involving respectively (a) the intermediate zone of the cerebellar hemisphere, and (b) the large lateral zone of this hemisphere. Let us discuss each of these separately.

CEREBELLAR FEEDBACK CONTROL OF THE DISTAL LIMBS THROUGH THE INTERMEDIATE CEREBELLAR ZONE AND INTERPOSITUS NUCLEUS

The intermediate zone of each cerebellar hemisphere receives information from two sources: (1) direct information from the motor cortex, and (2) feedback information from the peripheral parts of the body, especially from the distal portions of the limbs. After the cerebellum has integrated this information, output signals are then transmitted mainly to the cerebral cortex through relays in the interpositus nucleus and the thalamus. In addition, signals pass directly from the interpositus nucleus to the magnocellular portion (the lower portion) of the red nucleus that gives rise to the

rubrospinal tract. The rubrospinal tract in turn innervates especially those portions of the spinal cord gray matter that control the distal parts of the limbs, particularly the hands and fingers.

It is believed that this part of the cerebellar motor control system provides smooth, coordinate movements of the agonist and antagonist muscles of the distal limbs for the performance of acute purposeful intricate movements. The cerebellum seems to compare the "intentions" of the higher levels of the motor control system, as transmitted to the intermediate cerebellar hemisphere through the corticopontocerebellar tract, with the "performance" by the respective parts of the body as transmitted back to the cerebellum from the periphery. In fact, the ventral spinothalamic tract even transmits back to the cerebellum a "copy" of the actual motor control signals that reach the anterior motor neurons, and this too is integrated with the signals arriving from the muscle spindles and other proprioceptor sensory organs.

Ordinarily, during rapid movements, the motor cortex transmits far more impulses than are needed to perform each intended movement, and the cerebellum therefore must act to inhibit the motor cortex at the appropriate time after the muscle has begun to move. The cerebellum is believed to assess the rate of movement and calculate the length of time that will be required to reech the point of intention. Then appropriate inhibitory impulses are transmitted to the motor cortex to inhibit the agonist muscle and to excite the antagonist muscle. In this way, appropriate "brakes" are applied to stop the movement at the precise point of intention.

Thus, when e rapid movement is mede toward a point of intention, tbe agonist muscle contracts strongly throughout the early course of movement. Then, suddenly, shortly before the point of intention is reached, the agonist muscle becomes completely inhibited while the antagonist muscle becomes strongly excited. Furthermore, the point at which this reversal of excitation occurs depends on the rate of movement and on the previously learned knowledge of the inertia of the system. The faster the movement and the greater the inertia, the earlier the reversal point appears in the course of movement.

Since all these events transpire much too rapidly for the motor cortex to reverse the excitation "voluntarily," it is evident that the excitation of the antagonist muscle toward the end of a movement is an entirely automatic and subconscious function and is not a "willed" contraction of the same nature as the original contraction of the agonist muscle. We shall see later that in patients with serious cerebellar damage, excitation of the antagonist muscles does not occur at the appropriate time but instead always too late. Therefore, it is almost certain that one of the major functions of the cerebellum is automatic excitation of antagonist muscles at the end of a movement while at the same time inhibiting agonist muscles that have started the movement.

The "Damping" Function of the Cerebellum. One of the byproducts of the cerebellar feedback mechanism is its ability to "damp" muscular movements. To explain the meaning of "damping" we must first point out that essentially all movements of the body are "pendular." For instance, when an arm is moved, momentum develops, and the momentum must be overcome before the movement can be stopped. And, because of the momentum, all pendular movements have a tendency to overshoot. If overshooting does occur in a person whose cerebellum has been destroyed, the conscious centers of the cerebrum eventually recognize this and initiate a movement in the opposite direction to bring the arm to its intended position. But again the arm, by virtue of its momentum, overshoots, and appropriate corrective signals must again be instituted. Thus, the arm oscillates back and forth past its intended point for several cycles before it finally fixes on its mark. This effect is called an *action tremor*, or *intention tremor*.

However, if the cerebellum is intect, appropriate subconscious signals stop the movement precisely at the intended point, thereby preventing the overshoot and also the tremor. This is the besic characteristic of a damping system. All servocontrol systems regulating penduler elements tbet have inertia must have damping circuits built into the servomechanisms In the motor control system of our central nervous system, the cerebellum seems to provide much of this damping function.

Cerebellar Control of Ballistic Movements. Many rapid movements of the body, such as the movements of the fingers in typing, occur so rapidly that it is not possible to receive feedback information either from the periphery to the cerebellum or from the cerebellum back to the motor cortex before the movements are over. These movements are called *ballistic movements*, meaning that the entire movement is preplanned and is set into motion to go a specific distance and then to stop. Another important example is the saccadic movements of the eyes, in which the eyes jump from one position to the next when reading or when looking at successive points along a road when a person is moving in a car.

Much can be understood about tbe function of the cerebellum by studying the changes that occur in the ballistic movements when the cerebellum is removed. Three major changes occur: (1) the movements are slow to begin, (2) the force development is weak, and (3) the movements are slow to turn off. Therefore, it becomes very difficult to perform the very rapid ballistic movements. Furthermore, it is almost impossible to control how far the movement will go because of the difficulty of turn-

ing the movement off once it is begun. Thus, in the absence of the cerebellar circuit, the motor cortex has to think very hard to turn ballistic movements on and again has to think hard and take extra time to turn the movement off. Thus, the automatism of ballistic movements is lost.

But, how does the cerebellum function in the control of ballistic movements? We do not know the answer to this. The supposition is. When the motor cortex first initiates the movement, it immediately sends signals to the cerebellum at the same time. The first effect of the signals is to excite the deep cerebellar nuclei, and these immediately send an excitatory signal back to the motor cortex, red nucleus, or other motor nuclei to reinforce strongly the onset of the ballistic movement. A few milliseconds later, the signal entering the cerebellum will have had time to go through the delay circuits of the cerebellar cortex and to return by way of the Purkinje cells to the deep cerebellar nuclei, but this time inhibiting these rather than exciting them. Therefore, after this given delay time, this automatic delayed inhibitory signal presumably stops the ballistic movement by turning off the agonist muscle and, because of reciprocal innervation, turning on the antagonist at the same time.

If the student will consider once again the circuitry of the cerebellum as described earlier in the chapter, she or he will see that it is beautifully organized to perform this biphasic, first excitatory and then delayed inhibitory, function that is required for ballistic movements. The student will also see that the time delay circuits of the cerebellar cortex almost undoubtedly are fundamental to this particular ability of the cerebellum.

FUNCTION OF THE LARGE LATERAL ZONE OF THE CEREBELLAR HEMISPHERE—THE "PLANNING" AND "TIMING" FUNCTIONS

In human beings, the lateral zones of the two cerebellar hemispheres have become very highly developed and greatly enlarged, along with the human capability to perform intricate movements with the hands and fingers and along with the ability to speak. Yet, strangely enough, these large lateral portions of the cerebellar hemispheres have no direct input of information from the peripheral parts of the body. Also, almost all the communication between these lateral cerebellar areas and the cortex is not with the primary motor cortex itself but instead with the premotor area and primary and association somatic sensory areas. Even so, destruction of the lateral portions of the cerebellar hemispheres along with their deep nuclei, the dentate nuclei, can lead to extreme incoordination of the purposeful movements of the hands, fingers, feet, and speech apparatus. This has been hard to understand because of lack of direct communication between this part of the cerebellum and the primary motor cortex. However, recent experimental studies suggest that these portions of the cerebellum are concerned with two important aspects of motor control: (1) the "planning" of sequential movements, and (2) the "timing" of the sequential movements.

The Planning Function. The planning of sequential movements seems to be related to the fact that the lateral hemispheres communicate with the premotor and sensory portions of the cerebral cortex and that there is in turn two-way communication between these areas and corresponding areas of the basal ganglia. It seems that the "plan" of the sequential movements is transmitted from the sensory and premotor areas of the cortex to the lateral zones of the cerebellar hemispheres and two-way traffic between the cerebellum and the cortex is necessary to provide appropriate transition from one movement to the next. An exceedingly interesting observation that supports this view is that many of the neurons in the dentate nuclei display the activity pattern of the next movement at the same time that the present movement is occurring. Thus, the lateral hemispheres appear to be involved not with what is happening at a given moment but instead with *what will be happening during the next sequential movement.*

To summarize, one of the most important features of normal motor function is one's ability to progress smoothly from one movement to the next in orderly succession. In the absence of the cerebellar hemispheres this capability is seriously disturbed, especially for rapid movements.

The Timing Function. Another important function of the lateral cerebellar hemispheres is to provide appropriate timing for each movement. In the absence of these lateral areas, one loses the subconscious ability to predict ahead of time how far the different parts of the body will move in a given time. And without this timing capability, the person becomes unable to control the beginning of the next movement. As a result, the succeeding movement may begin much too early or much too late. Therefore, cerebellar lesions cause complex movements, such as those required for writing, running, or even talking, to become completely incoordinate, lacking completely in the ability to progress in an orderly sequence from one movement to the next. Such cerebellar lesions are said to cause *failure of smooth progression of movements.*

Extramotor Predictive Functions of the Cerebellum. The cerebellum also plays a role in predicting other events besides movements of the body. For instance, the rates of progression of both auditory and visual phenomena can be predicted. As an example, a person can predict from the changing visual scene how rapidly he or she is approaching an object. A striking experiment that

demonstrates the importance of the cerebellum in this ability is the effect of removing the "bead" portion of the cerebellum in monkeys. Such a monkey occasionally charges the wall of a corridor and literally bashes its brains out because it is unable to predict when it will reach the wall.

Unfortunately, we are only now beginning to learn about these extramotor predictive functions of the cerebellum. It is quite possible that the cerebellum provides a "time base," perhaps utilizing time-delay circuits, against which signals from other parts of the central nervous system can be compared. It is often stated that the cerebellum is especially important in interpreting *spatiotemporal relationships* in sensory information.

CLINICAL ABNORMALITIES OF THE CEREBELLUM

Essentially all lesions or malfunctions of the cerebellum cause abnormal motor movements on the *same side*—the *ipsilateral side*—of the body as the side of the malfunction in the cerebellum. The reason for this is that almost all tracts feeding sensory information into the cerebellum, especially the dorsal spinocerebellar tract, are uncrossed, whereas all tracts in both directions between the cerebral cortex and the cerebellum cross to the opposite side in the same manner that the pyramidal tract crosses to the opposite side of the peripheral body.

A second important feature of clinical cerebellar abnormalities is that destruction of small portions of the cerebellar cortex rarely causes a detectable abnormality in motor function. In fact, several months after as much as half the cerebellar cortex has been removed, if the deep cerebellar nuclei are not removed along with the cortex the motor functions of an animal appear to be almost entirely normal as long as the animal performs all movements slowly. Thus, the remaining portions of the motor control system are capable of compensating tremendously for loss of parts of the cerebellum.

To cause serious and continuous dysfunction of the cerebellum, the cerebellar lesion must usually involve the deep cerebellar nuclei—the *dentate*, *interpositus*, and *fastigial nuclei*—as well as the cerebellar cortex.

Dysmetria and Ataxia. Two of the most important symptoms of cerebellar disease are dysmetria and ataxia. It was pointed out earlier that in the absence of the cerebellum the subconscious motor control system cannot predict ahead of time how far movements will go. Therefore, the movements ordinarily overshoot their intended mark, though sometimes they are short of the mark. This effect is called *dysmetria*, and it results in incoordinate movements that are called *ataxia*.

Dysmetria and ataxia can also result from lesions in the spinocerebellar tracts, for the feedback information from the moving parts of the body is essential for accurate control of the movements.

Past Pointing. Past pointing means that in the absence of the cerebellum a person ordinarily moves the hand or some other moving part of the body considerably beyond the point of intention. This probably results from the following effect: The motor cortex normally transmits more impulses to the muscles to perform a given motor function than are actually needed. The cerebellum automatically corrects this by inhibiting the movement after it has begun. However, if the cerebellum is not available to cause this inhibition, the movement ordinarily goes beyond the intended point. Therefore, past pointing is actually a manifestation of dysmetria.

Failure of Progression. *Dysdiadochokinesia.* When the motor control system fails to predict ahead of time where the different parts of the body will be at a given time, it temporarily "loses" the parts during rapid motor movements. As a result, the succeeding movement may begin much too early or much too late so that no orderly "progression of movement" can occur. One can demonstrate this readily by having a patient with cerebellar damage turn one hand upward and downward at a rapid rate. The patient rapidly "loses" the hand and does not know its position during any portion of the movement. As a result, a series of jumbled movements occurs instead of the normal coordinate upward and downward motions. This is called *dysdiadochokinesia*.

Dysarthria. Another instance in which failure of progression occurs is in talking, for the formation of words depends on rapid and orderly succession of individual muscular movements in the larynx, mouth, and respiratory system. Lack of coordination between these and inability to predict either the intensity of the sound or the duration of each successive sound cause jumbled vocalization, with some syllables loud, some weak, some held long, some held for short intervals, and resultant speech that is almost completely unintelligible. This is called *dysarthria*.

Intention Tremor. When a person who has lost the cerebellum performs a voluntary act, the muscular movements are jerky; this reaction is called an *intention tremor* or an *action tremor*, and it results from failure of the cerebellar system to damp the motor movements.

Cerebellar Nystagmus. Cerebellar nystagmus is a tremor of the eyeballs that occurs usually when one attempts to fixate the eyes on a scene to one side of the head. This off-center type of fixation results in rapid, tremulous movements of the eyes rather than a steady fixation, and it is probably another manifestation of the failure of damping by the cerebellum. It occurs especially when the flocculonodular lobes are damaged; in this instance it is associated with loss of equilibrium, presumably because of dysfunction of the pathways through the cerebellum from the semicircular canals.

Rebound. If a person with cerebellar disease is asked to pull upward strongly on an arm while the physician holds it back at first and then lets go, the arm will fly back until it strikes the face instead of being automatically stopped. This is called *rebound*, and it results from *loss of the cerebellar component of the stretch reflex*. That is, the normal cerebellum ordinarily instantaneously and powerfully sensitizes this spinal cord reflex mechanism whenever a portion of the body begins to move unexpectedly in an unwilled direction. But, without the cerebellum, strong activation of the antagonist muscles fails to occur, thus allowing overmovement of the limb.

Hypotonia. Loss of the deep cerebellar nuclei, particularly the dentate and interpositus nuclei, causes decreased tone of the peripheral musculature on the side of the lesion, though after several months the motor cortex usually compensates for this by an increase in its intrinsic activity. The hypotonia results from loss of facilitation of the motor cortex and brain stem nuclei by the tonic discharge of the deep nuclei.

SENSORY FEEDBACK CONTROL OF MOTOR FUNCTIONS

THE "SENSORY ENGRAM" FOR MOTOR CONTROL

When large portions of the somatic sensory cortex are removed or damaged, a person or an animal loses the ability to perform many complex motor activities previously learned The reason for this is believed to be that, when a person learns to perform a motor function, he or she experiences in the somatic sensory areas the effects of the motor movement each time it is performed, and "memories" of the different patterns of movements are recorded. These are called *sensory engrams of the motor movements.* Once a sensory engram has been established in the sensory cortex—that is, once the movement has been learned—the person then uses this sensory engram as a guide for the motor system of the brain to follow in reproducing the same pattern of movement. For instance, as a person learns to cut out a paper doll with scissors, the movements involved in this process cause a particular sequential pattern of sensory signals to pass to the somatic sensory area. Once this pattern has been "learned" by the sensory cortex, the memory engram of the pattern can be used to activate the motor system to perform the same sequential pattern whenever it is required—even when the person is blindfolded.

To do this, the sensory signals from the fingers, hands, and arms are compared with the engram, and if the two do not match each other, the difference, called the "error," supposedly initiates additional motor signals that automatically activate appropriate muscles to bring the fingers, hands, and arms into the necessary sequential attitudes for performance of the task. Each successive portion of the engram presumably is projected according to a time sequence, and the motor control system automatically follows from one point to the next so that the fingers go through the precise motions necessary to duplicate the sensory engram

Thus, one can see that the motor system in this case acts as a *servomechanism*, for it is not the motor cortex itself that controls the pattern of activity to be accomplished. Instead, the pattern is located in the sensory part of the brain, and the motor system merely "follows" the pattern, which is the definition of a servomechanism. If ever the motor system fails to follow the pattern, sensory signals are fed back to the sensory cortex to apprise the sensorium of this failure, and appropriate corrective signals are transmitted to the muscles

Other sensory signals besides somesthetic signals are also involved in motor control, particularly visual signals. However, these other sensory systems are often slower to recognize error than is the somatic system. Therefore, when the sensory engram depends on visual feedback for control purposes, the movements are usually considerably slowed in comparison with those that depend on somatic feedback.

An extremely interesting experiment that demonstrates the importance of the sensory engram for control of motor movements is one in which a monkey has been trained to perform some complex but slow-moving activity, and then various portions of its cortex are removed. Removal of small portions of the motor cortex that control the muscles normally used for the activity does not prevent the monkey from performing the activity. Instead it automatically uses other muscles in place of the paralyzed ones to perform the same activity. On the other hand, if the corresponding somatic sensory cortex is removed while the motor cortex is left intact, the monkey loses all ability to perform the activity Thus, this experiment demonstrates that the motor system acts automatically as a servomechanism to use whatever muscles are available to follow the pattern of the sensory engram, and if some muscles are missing, other muscles are substituted automatically. The experiment also demonstrates forcefully that the somatic sensory cortex is essential for performance of most complex types of "learned" motor activities.

ESTABLISHMENT OF RAPID MOTOR PATTERNS

Many motor activities are performed so rapidly that there is insufficient time for sensory feedback signals to control these activities. For instance, the movements of the fingers during typing occur much too rapidly for somatic sensory signals to be transmitted either to the somatic sensory cortex or even directly to the motor cortex and for these then to control each discrete movement. It is believed that the patterns for control of these rapid, coordinate muscular movements are established in the motor system itself, probably involving complex circuitry in the primary motor cortex, in the premotor cortex, the basal ganglia, and even the cerebellum Indeed, lesions in any of these areas can destroy one's ability to perform rapid coordinated muscular contractions, such as those required during the acts of typing, talking, or writing by hand.

Role of Sensory Feedback During Establishment of the Rapid Motor Patterns. Even a highly skilled motor activity can be performed the very first time if it is performed extremely slowly—slowly enough for sensory feedback to guide the movements through each step. However, to be really useful, many skilled motor activities must be performed rapidly. This probably is achieved by successive performance of the same skilled activity until finally an engram for the skilled activity is

aid down in the motor control areas of the cortex is well as in the sensory system. This *motor ngram* causes a precise set of muscles to go hrough a specific sequence of movements required o perform the skilled activity. Therefore, such an ngram is called a *pattern of skilled motor function*, and the motor areas are primarily concerned with this.

After a person has performed a skilled activity many times, the motor pattern of this activity can thereafter cause the hand or arm or other part of the body to go through the same pattern of activity again and again, now entirely *without* sensory feedback control. However, even though sensory feedback control is no longer present, the sensory system still determines whether or not the act has been performed correctly. This determination is made in retrospect rather than while the act is being performed. If the pattern has not been performed correctly, information from the sensory system can help correct the pattern the next time it is performed.

Thus, eventually, hundreds of patterns of different coordinate movements are believed to be laid down in the motor system, and these can be called upon one at a time in different sequential orders to perform literally thousands of complex motor activities.

An interesting experiment that demonstrates the applicability of these theoretical methods of muscular control is one in which the eyes are made to "follow" an object that moves around and around in a circle. At first, the eyes can follow the object only when it moves around the circle slowly, and even then the movements of the eyes are extremely jerky. Thus, sensory feedback is being utilized to control the eye movements for following the object. However, after a few seconds, the eyes begin to follow the moving object rather faithfully, and the rapidity of movement around the circle can be increased to many times per second, and still, the eyes continue to follow the object. Sensory feedback control of each stage of the eye movements at these rapid rates would be completely impossible. Therefore, by this time, the eyes have developed a pattern of movement that is not dependent upon step-by-step sensory feedback. Nevertheless, if the eyes should fail to follow the object around the circle, the sensory system would immediately become aware of this and presumably could make corrections in the pattern of movement.

INITIATION OF VOLUNTARY MOTOR ACTIVITY

Because of the spectacular properties of the primary motor cortex (the area pyramidalis) and of the instantaneous muscle contractions that can be achieved by stimulating this area, it has become customary to think that the initial brain signals that elicit voluntary muscle contractions begin in the primary motor cortex. However, this almost certainly is far from the truth. Indeed, experiments have shown that the cerebellum and the basal ganglia are activated at olmost exactly the same time as or even before the motor cortex is activated. Furthermore, there is no known mechanism by which the motor cortex can conceive the entire sequential pattern that is to be achieved by the motor movements.

Therefore, we are left with an unanswered question: What is the locus of the initiation of voluntary motor activity? A partial answer to this comes from the fact that neuronal activity occurs in sensory areas of the brain as long as a second before a voluntary motor activity occurs. Also, neuronal activity begins in the premotor areas of the cortex and in some areas of the basal ganglia many milliseconds before motor activity occurs in the motor cortex. Therefore, it is beginning to be believed that *cerebrotion* occurring in these "integrative" portions of the brain, operating in association with the cerebellum, conceives and plans the complex sequence of movements that is to be executed. Only after the "plan" has been established is the primary motor system involving the primary motor cortex set into action to causa tha sequential events In the following chapter, we shall discuss the planning and execution of speech, which is an excellent illustration of these principles of voluntary motor control.

REFERENCES

Allen, G I, and Tsukahara, N Cerebrocerebellar communication systems *Physiol Rev*, *54* 957, 1974

Armstrong, D M The mammalian cerebellum and its contribution to movement control *In* Porter, R. (ed) International Review of Physiology Neurophysiology III Vol 17 Baltimore, University Park Press, 1978, p 239

Asanuma, H Cerebral cortical control of movement *Physiologist*, *16* 143, 1973

Asanuma, H Recent developments in the study of the columnar arrangement of neurons within the motor cortex *Physiol Rev*, *55* 143, 1975

Asanuma, H The pyramidal tract. *In* Orooks, V B (ed.) Handbook of Physiology Sec 1, Vol II Oethesda, American Physiological Society, 1981, p 703

Dloedel, J R., and Courville, J Cerebellar afferent systems *In* Brooks, V B (ed) Handbook of Physiology Sec 1, Vol II. Bethesda, American Physiological Society, 1981, p 735

Brooks, V B, and Stuney, S D, Jr Motor mechanisms The role of the pyramidal system in motor control *Annu Rev Physiol*, *33*.337, 1971

Brooks, V B, and Thach, W T Cerebellar control of posture and movement *In* Brooks, V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 877

Cooper, I. S, *et al* (eds) The Cerebellum, Epilepsy, and Behavior New York, Plenum Press, 1974

Crowley, W J., Neural control of skeletal muscle *In* Frohlich, E D. (ed) Pathophysiology, 2nd Ed Philadelphia, J. B. Lippincott Co, 1976, p 735

Deamedt, J E (ed) Cerebral Motor Control in Man Long Loop Mechanisms New York, S Karger, 1978

Eccles, J C The Understanding of the Brain. New York, McGraw Hill, 1973

Evarts, E V Brain mechanisms in movement *Sci. Am*, *229* 96, 1973

Evarts, E V. Brain mechanisms of movement. *Sci. Am*, *241*(3) 164, 1979

Evarts, E V.: Role of motor cortex in voluntary movements in primates *In* Brooks, V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 1083
Evarts, E V, and Thach, W T Motor mechanisms of the CNS Cerebrocerebellar interrelations *Annu. Rev Physiol* 31 451 1969
Fuster, J M.. Prefrontal cortex in motor control *In* Brooks, V. B (ed) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 1149
Gallistel, C R. The Organization of Action A New Synthesis New York, Halsted press, 1979
Granit, R · The Basis of Motor Control New York, Academic Press, 1970
Grillner, S. Locomotion in vertebrates Central mechanisms and reflex interaction *Physiol Rev* 55 247, 1975
Keele, S W Behavioral analysis of movement *In* Brooks, V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda American Physiological Society, 1981, p 1391
Kuypers, H G J M Anatomy of the descending pathways *In* Brooks, V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda American Physiological Society, 1981, p 597
Llinas, R Eighteenth Bowditch lecture Motor aspects of cerebellar control *Physiologist*, 17 19, 1974
Llinas, R Electrophysiology of the cerebellar networks *In* Brooks V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 831
Massion, J, and Sasaki, K (eds) Cerebro-Cerebellar Interactions New York, Elsevier-North Holland 1979
McCloskey, D I Corollary discharges Motor commands and perception *In* Brooks, V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda American Physiological Society, 1981, p 1415
McHenry, L C., Jr · Cerebral Circulation and Stroke St Louis, W H Green, 1978
O Connell, A L., and Gardner, E B Understanding the Scientific Bases of Human Movement Baltimore Williams & Wilkins, 1972
Orlovsky, G N, and Shik, M L.· Control of locomotion A neurophysiological analysis of the rat locomotor system. *Int. Rev Physiol* 10 281, 1976
Paillard J The patterning of skilled movements *In* Magoun, H (ed) Handbook of Physiology Sec 1, Vol. 3, Baltimore, Williams Wilkins 1960, p 1679
Pearson, K The control of walking *Sci. Am*, 235 6·72, 1976
Penfield, W, and Rasmussen, T · The Cerebral Cortex of Man New York, The Macmillan Co, 1950
Porter, R Functions of the mammalian cerebral cortex in movement *Prog Neurobiol* 1.3, 1973
Porter, R The neurophysiology of movement performance *In* M International Review of Science Physiology, Vol 3 Baltimore, University Park Press 1974, p 151
Porter, R Influences of movement detectors on pyramidal tract neurons in primates *Annu Rev Physiol*, 38·121, 1976
Porter, R Internal organization of the motor cortex for input-output arrangements *In* Brooks, V B (ed) Handbook of Physiology Sec Vol II Bethesda, American Physiological Society, 1981, p 1063
Poulton, E C Human manual control *In* Brooks, V B (ed) Handbook of Physiology Sec. 1, Vol II Bethesda, American Physiological Society, 1981, p 1337
Shik, M L., and Orlovsky, G N Neurophysiology of locomotor automatism *Physiol Rev*, 56 465, 1976
Stein, P S G Motor systems with specific reference to the control locomotion *Annu Rev Neurosci*, 1 61, 1978
Wiesendanger M Organization of secondary motor areas of cerebral cortex *In* Brooks, V B (ed) Handbook of Physiology. Sec 1, Vol. Bethesda American Physiological Society, 1981, p 1121
Wiesendanger, M, and Miles, T S Ascending pathway of low threshold muscle afferents to the cerebral cortex and its possible role in motor control *Physiol Rev* 62 1234, 1982
Wilson D M The flight-control system of the locust. Sci Am 218 (5)·83 1968

54

The Cerebral Cortex and Intellectual Functions of the Brain

It is ironic that we know least about the mech-
ıisms of the cerebral cortex of almost all parts of
e brain, even though it is by far tho largest
ırtion of the nervous system. Yet, we do know
.e effects of destruction or of specific stimulation
various portions of the cortex, and still more is
ıown from electrical recordings from the cortex
from the surface of the scalp. In the early part
the present chapter the facts known about cor-
:al functions are discussed and then some basic
ıeories of the neuronal mechanisms involved in
ıought processes, memory, analysis of sensory
ıformation, and so forth are presented briefly.

‹YSIOLOGIC ANATOMY F THE CEREBRAL CORTEX

The functional part of the cerebral cortex is composed
ainly of a thin layer of neurons 2 to 5 millimeters in
ıickness, covering the surface of all the convolutions
the cerebrum and having a total area of about one
ıarter square meter. The total cerebral cortex probably
ıntains 50 to 100 billion neurons.

Figure 54–1 illustrates the typical structure of the
ırebral cortex, showing successive layers of different
ypes of cells. Most of the cells are of three types.
anular (also called *stellate*), *fusiform*, and *pyramidal*,
ıe latter named for their characteristic pyramidal
ıape. To the right in Figure 54–1 is illustrated the
ypical organization of nerve fibers within the different
ıyers of the cortex. Note particularly the large number
horizontal fibers extending between adjacent areas of
ıe cortex, but note also the *vertical fibers* that extend
) and from the cortex to lower areas of the brain and
) the spinal cord or to distant regions of the cerebral
ırtex through long association bundles of fibers.

Neurohistologists have divided the cerebral cortex
ıto almost 100 different areas, which have slightly
ıfferent architectural characteristics. Yet in all these
ıfferent areas except the hippocampal region there still
ersist representations of all the six major layers of the
ırtex. To the untrained histologist, only five major
rchitectural types of cortex can be distinguished as
follows Type 1 contains large numbers of pyramidal
cells with few granular cells and, therefore, is often
called the *agranular cortex*. At the other extreme, type
5 contains almost no pyramidal cells but is filled with
closely packed granule cells and is called the *granular
cortex* Types 2, 3, and 4 have intermediate character-
istics containing gradated proportions of pyramidal and
granular cells.

The agranular cortex, which contains large numbers
of pyramidal cells, is characteristic of the motor areas
of the cerebral cortex, whereas the granular cortex,
containing almost no pyramidal cells, is characteristic

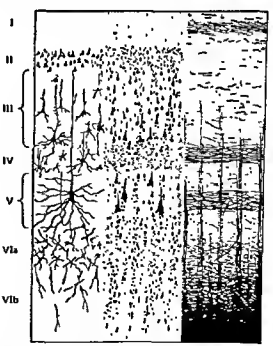


Figure 54–1. Structure of the cerebral cortex, illustrating *I*, molecular layer, *II*, external granular layer; *III*, layer of pyramidal cells, *IV*, internal granular layer; *V*, large pyramidal cell layer; *VI*, layer of fusiform or polymorphic cells. (from Ranson and Clark (after Brodmann) Anatomy of the Nervous System. Philadelphia, W. B. Saunders Company, 1959.)

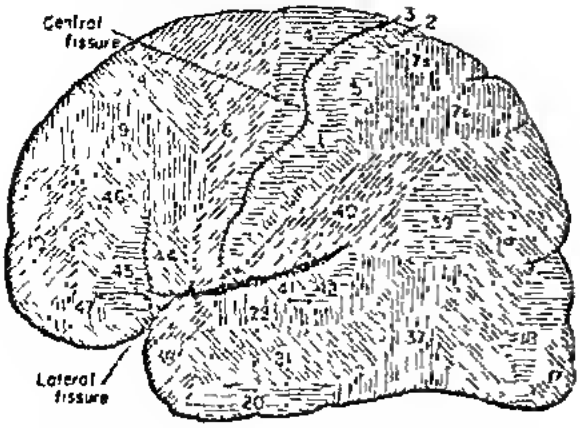


Figure 54–2. Structurally distinct areas of the human cerebral cortex. (From Everett Functional Neuroanatomy 5th ed Philadelphia, Lea & Febiger, 1965 Modified from Brodmann)

of the primary sensory areas The other three intermediate types of cortex are characteristic of the association areas between the primary sensory and motor regions The pyramidal cells give rise to the long axons that transmit motor signals to the basal ganglia, the brain stem and cerebellum, and the spinal cord The granule cells have short axons that transmit multiple signals from neuron to neuron within the neural network of the cortex itself.

Figure 54–2 shows a map of cortical areas classified on the basis of histological studies by Brodmann This classification provides a basis for discussion of functional areas of the brain, even though different architectural types of cortex often have similar functions.

Anatomical Relationship of the Cerebral Cortex to the Thalamus and Other Lower Centers. All areas of the cerebral cortex have direct afferent and efferent connections with the thalamus Figure 54–3 shows the areas of the cerebral cortex connected with specific parts of the thalamus These connections are in *two* directions, both from the thalamus to the cortex and then from the cortex back to essentially the same area of the thalamus Furthermore, when the thalamic connections are cut, the functions of the corresponding cortical area become entirely or almost entirely abrogated Therefore, the cortex operates in close association with the thalamus and can almost be considered both anatomically and functionally to be a large outgrowth of the thalamus, for this reason the thalamus and the cortex together are often called the *thalamocortical system*, as we shall discuss further in the following chapter. Also, all pathways from the sensory organs to the cortex pass through the thalamus, with the single exception of the sensory pathways of the olfactory tract.

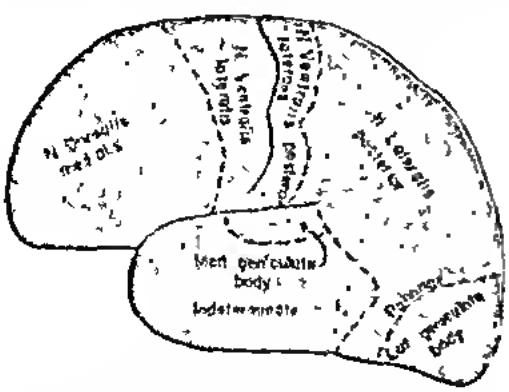


Figure 54–3. Areas of the cerebral cortex that connect with specific portions of the thalamus. (Modified from Elliott Textbook of the Nervous System. J B Lippincott Co)

FUNCTIONS OF CERTAIN SPECIFIC CORTICAL AREAS

Studies in human beings by neurosurgeons have shown that some specific functions are localized to certain general areas of the cerebral cortex. Figure 54–4 gives a map of some of these areas as determined by Penfield and Rasmussen from direct electrical stimulation of the cortex or by neurological examination of patients after portions of the cortex had been removed The lightly shaded areas are *primary sensory areas*, while the darkly shaded area is the *primary motor area* (also called *voluntary motor area*) from which muscular movements can be elicited with relatively weak electrical stimuli. These primary sensory and motor areas have highly specific functions as we have discussed in previous chapters, whereas other areas of the cortex perform more general functions that we call association or cerebration

SPECIFIC FUNCTIONS OF THE PRIMARY SENSORY AREAS

The primary sensory areas all have certain functions in common. For instance, the somatic sensory areas, visual sensory areas, and auditory sensory areas all have spatial localizations of signals from the peripheral receptors (which are discussed in detail in Chapters 49, 60, and 61).

Electrical stimulation of the *primary somatic sensory area* in the parietal lobe in awake patients elicits relatively uncomplicated sensations. For instance, in the somatic sensory area the patient

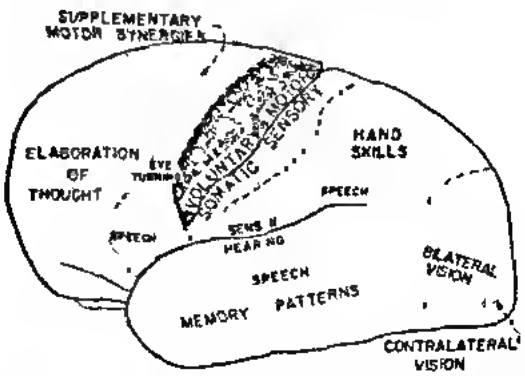


Figure 54–4. Functional areas of the human cerebral cortex as determined by electrical stimulation of the cortex during neurosurgical operations and by neurological examinations of patients with destroyed cortical regions (From Penfield and Rasmussen The Cerebral Cortex of Man A Clinical Study of Localization of Function. New York, Macmillan Company, 1968)

expresses feelings of tingling in the skin, numbness, mild "electric" feeling, or, rarely, mild degrees of temperature sensations. And these sensations are localized to discrete areas of the body in accord with the spatial representation in the somatic sensory cortex, as described in Chapter 49. Therefore, it is believed that the primary somatic sensory cortex analyzes only the simple aspects of sensations and that analysis of intricate patterns of sensory experience requires also adjacent parts of the parietal lobes called the *somatic sensory association area.*

Electrical stimulation of the *primary visual cortex* in the occipital lobe causes the person to see flashes of light, bright lines, colors, or other simple visions. Here again, the visual images are localized to specific regions of the visual fields in accord with the portion of the primary visual cortex stimulated, as described in Chapter 60. But the visual cortex alone is not capable of complete analysis of complicated visual patterns; for this, the visual cortex must operate in association with adjacent regions of the occipital, temporal, and parietal cortex, the *visual association areas.*

Electrical stimulation of the *primary auditory cortex* in the temporal lobe causes a person to hear a simple sound which may be weak or loud, have low or high frequency, or have other uncomplicated characteristics, such as a squeak or even an undulation. But never are words or any other fully intelligible sound heard. Thus, the primary auditory cortex, like the other primary sensory areas, can detect the individual elements of auditory experience but cannot analyze complicated sounds. Therefore, the primary auditory cortex alone is not sufficient to give one even the usual auditory experiences; these can be achieved, however, when the primary area operates together with the *auditory association areas* in adjacent regions of the temporal lobes.

Despite the inability of the primary sensory areas to analyze the incoming sensations fully, when these primary areas are destroyed the ability of the person to utilize the respective sensations usually suffers drastically. For instance, loss of the primary visual cortex in one occipital lobe causes a person to become blind in the ipsilateral halves of both retinae, and loss of the primary visual cortices in both hemispheres causes total blindness. Likewise, loss of both primary auditory cortices causes almost total deafness. Loss of the primary somatic sensory areas in the postcentral gyri causes *depression* of somatic sensory sensations—though not total loss—presumably because of additional thalamic and cortical representation of these sensations in other cortical areas: somatic sensory area II and the motor cortex, for example. (In animals far down the phylogenetic scale, loss of the visual and auditory cortices may have little effect on vision and hearing, and even an anencephalic human infant having no brain above the mesencephalon detects some visual cues in the absence of the visual cortex, for such a child can observe a moving object and even follow it by movement of the eyes and head.)

Therefore, we can summarize the functions of the primary sensory areas of the human cerebral cortex in the following way: The lower centers of the brain relay a large part of the sensory signals to the cerebral cortex for analysis. In turn, the primary sensory areas transmit the results of their analyses, albeit still simple analyses, back to the lower centers and to other regions of the cerebral cortex, as is discussed later in the chapter.

THE SENSORY ASSOCIATION AREAS

Around the borders of the primary sensory areas are regions called *sensory association areas* or *secondary sensory areas.* In general, these areas extend 1 to 5 centimeters in one or more directions from the primary sensory areas; each time a primary area receives a sensory signal, secondary signals spread, after a delay of a few milliseconds, into the respective association area as well. Part of this spread occurs directly from the primary area through subcortical fiber tracts, but part also occurs in the thalamus, beginning in the sensory relay nuclei, passing next to corresponding *thalamic association areas,* and then traveling to the association cortex

The general function of the sensory association areas is to provide a higher level of interpretation of the sensory experiences. The general areas for the interpretative functions for somatic, visual, and auditory experiences are illustrated in Figure 54–5

Destruction of the *sensory association area* greatly reduces the capability of the brain to analyze different characteristics of sensory experiences For instance, damage in the temporal lobe below and behind the primary auditory area in the "dominant hemisphere" of the brain often causes a person to lose the ability to understand

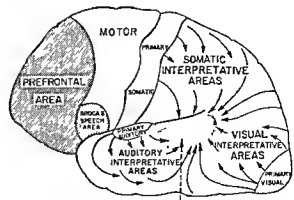


Figure 54–5. Organization of the somatic, auditory, and visual association areas into a general mechanism for interpretation of sensory experience. All these feed also into Wernicke's area located in the posterosuperior portion of the temporal lobe Note also the prefrontal area, and Broca's speech area.

words or other auditory experiences even though they are heard.

Likewise, destruction of the *visual association area* in Brodmann's areas 18 and 19 of the occipital lobe in the dominant hemisphere (see Figure 54–2), or the presence of a brain tumor or other lesion in these areas, does not cause blindness or prevent normal activation of the primary visual cortex but does greatly reduce the person's ability to interpret what is seen. Such a person often loses the ability to recognize the meanings of words, a condition that is called *word blindness* or *dyslexia*.

Finally, destruction of the *somatic sensory association area* in the parietal cortex posterior to primary somatic area I causes the person to lose spatial perception for location of the different parts of the body. In the case of the hand that has been "lost," the skills of the hand are greatly reduced. Thus, this area of the cortex seems to be necessary for interpretation of somatic sensory experiences.

The functions of the association areas are described in more detail in Chapter 49 for somatic, in Chapter 60 for visual, and in Chapter 61 for auditory experiences.

INTERPRETATIVE FUNCTION OF THE POSTERIOR SUPERIOR TEMPORAL LOBE—"WERNICKE'S AREA" (A GENERAL INTERPRETATIVE AREA)

The somatic, visual, and auditory association areas, which can actually be called "interpretative areas," all meet one another in the posterior part of the superior temporal lobe where the temporal, parietal, and occipital lobes all come together. This area of confluence of the different sensory interpretative areas is especially highly developed in the dominant side of the brain—the *left side* in right-handed persons—and it plays the greatest single role of any part of the cerebral cortex in the higher levels of brain function that we call *cerebration*. Therefore, this region has frequently been called by different names suggestive of the area having almost global importance: the *general interpretative area*, the *gnostic area*, the *knowing area*, the *tertiary association area*, and so forth. However, it is best known as *Wernicke's area* in honor of the neurologist who first described its special significance in intellectual processes.

Following severe damage in Wernicke's area, a person might hear perfectly well and even recognize different words but still might be unable to arrange these words into a coherent thought. Likewise, the person may be able to read words from the printed page but be unable to recognize the thought that is conveyed.

Electrical stimulation in Wernicke's area of the conscious patient occasionally causes a highly complex thought. This is particularly true when the stimulatory electrode is passed deep enough into the brain to approach the corresponding connecting areas of the thalamus. The types of thoughts that might be experienced include memories of complicated visual scenes that one might remember from childhood, auditory hallucinations such as a specific musical piece, or even a discourse by a specific person. For this reason it is believed that complicated memory patterns involving more than one sensory modality are stored at least partially in this area, or that activation of Wernicke's area can call forth the memories even if they may be stored elsewhere. This belief is in accord with the importance of Wernicke's area in interpretation of the complicated meanings of different sensory experiences.

The Angular Gyrus—Interpretation of Visual Information. The angular gyrus is the most inferior portion of the posterior parietal lobe, lying immediately behind Wernicke's area and fusing posteriorly into the visual areas of the occipital lobe as well. If this region is destroyed while Wernicke's area in the temporal lobe is still intact, the person can still interpret auditory experiences as usual, but the stream of visual experiences passing into Wernicke's area from the visual cortex is mainly blocked. Therefore, the person may be able to see words and even know they are words but, nevertheless, not be able to interpret their meanings. This is the condition called *dyslexia*, or *word blindness*.

Let us again emphasize the global importance of Wernicke's area for most intellectual functions of the brain. Loss of this area in an adult usually leads thereafter to a lifetime of almost demented existence.

The Concept of the Dominant Hemisphere

The general interpretative functions of Wernicke's area and of the angular gyrus, and also the functions of the speech and motor control areas, are usually much more highly developed in one cerebral hemisphere than in the other. Therefore, this hemisphere is called the *dominant hemisphere*. In about 95 per cent of all persons the left hemisphere is the dominant one. Wernicke's area of the brain is as much as 50 per cent larger in the left hemisphere than in the right in more than one half of newborn babies. Therefore, it is easy to understand why the left side of the brain might become dominant over the right side. However, if for some reason the dominant Wernicke's area is removed in early childhood, the opposite side of the brain can develop full dominant characteristics.

A theory that can explain the capability of one hemisphere to dominate the other hemisphere is the following:

The attention of the "mind" seems to be directed to one portion of the brain at a time. Presumably, because its size is usually larger at birth, the left temporal lobe normally begins to be used to a

reater extent than the right, and, thenceforth, ecause of the tendency to direct one's attention o the better developed region, the rate of learning n the cerebral hemisphere that gains the first tart increases rapidly while that in the opposite ide remains slight. Therefore, in the normal hunan being, ono side becomes dominant over the ther.

In about 95 per cent of all persons the left emporal lobe and angular gyrus become dominant, and in the remaining 5 per cent of the opulation either both sides develop simultaneously to have dual dominance, or, more rarely, the ight side alone becomes highly developed

Usually associated with the dominant temporal lobe and angular gyrus is dominance of certain portions of the somatic sensory cortex and motor cortex for control of voluntary motor functions. For nstance, as is discussed later in the chapter, the premotor speech area (Broca's area), located far laterally in the intermediate frontal lobe, is almost always dominant also on the left side of the brain. This speech area causes the formation of words by exciting simultaneously the laryngeal muscles, the respiratory muscles, and the muscles of the mouth. The motor areas for controlling the hands are dominant on the left side of the brain in about alne of ten persons, thus causing "right-handedaess" in most people.

Though the interpretative areas of the temporal lobe and angular gyrus, as well as many of the motor areas, are highly developed in only a single hemisphere, they are capable of receiving sensory informatioa from both hemispheres and are also capable of controlling motor activities in both hemispheres, utilizing mainly fiber pathways in the *corpus callosum* for communication between the two hemispheres. This unitary, cross-feeding organization prevents interference between the functions of the two sides of the brain; such interference, obviously, could create havoc with both thoughts and motor responses.

Role of Language in Function of Wernicke's Area and in Intellectual Functions

A major share of our sensory experience is converted into its language equivalent before being stored in the memory areas of the brain and before being processed for other intellectual purposes. For instance, when we read a book, we do not store the visual images of the printed words but, instead, store the words themselves in language form. Also, the information conveyed by the words is usually converted to language form before its meaning is discerned.

The sensory area of the dominant hemisphere *for* interpretation of language is Wernicke's area, and this is very closely associated with both the primary hearing area and the auditory association areas of the temporal lobe. This very close relationship probably results from the fact that the first introduction to language is by way of hearing. Later in life, when visual perception of language through the medium of reading develops, the visual information is then presumably channeled into the already developed language regions of the dominant temporal lobe.

Wernicke's Area in the Nondominant Hemisphere. When Wernicke's area in the dominant hemisphere is destroyed, the person normally loses almost all intellectual functions associated with language or verbal symbolism, such as ability to read, ability to perform mathematical operations, and even the ability to think through logical problems. However, many other types of interpretative capabilities, some of which utilize the temporal lobe and angular gyrus regions of the opposite hemisphere, are retained. Psychological studies in patients with damage to their nondominant hemispheres have suggested that this hemisphere may be especially important for understanding and interpreting music, nonverbal visual experiences (especially visual patterns), spatial relationships between the person and the surroundings, and probably also interpreting many somatic experiences related to use of the limbs and hands Thus, even though we speak of the "dominant" hemisphere, this dominance is primarily for language- or verbal symbolism-related intellectual functions; the opposite hemisphere is actually dominant for some other types of intelligence.

An Area for Recognition of Faces

An interesting type of brain abnormality called *prosophenosia* is the inability to recognize faces This occurs in persons who have extensive damage on the medial undersides of both occipital lobes and along the medioventral surfaces of the temporal lobes, as illustrated in Figure 54–6 Loss of these face recognition areas, *strangely enough*, results in very little other abnormality of brain function

Figure 54–6. Facial recognition areas located on the underside of the brain in the medial occipital and temporal lobes. (from Geschwind *Sci Am.*, *241* 180, 1979. © 1979 by Scientific American, Inc. All rights reserved.)

One wonders why so much of the cerebral cortex should be reserved for the simple task of face recognition. However, when it is remembered that most of our daily tasks involve associations with other people, one can see the importance of this intellectual function.

The occipital portion of this area is contiguous with the primary visual cortex, and the temporal portion is closely associated with the limbic system that has to do with emotions, brain activation, and control of one's behavioral response to the environment, as we shall discuss in Chapter 56

THE PREFRONTAL AREAS

The prefrontal areas are those portions of the frontal lobes that lie anterior to the motor regions, as shown in Figure 54–5 For years, this part of the cortex has been considered to be the locus of the higher intellect of the human being, principally because the main difference between the brain of monkeys and that of human beings is the great prominence of the human prefrontal areas. Yet efforts to show that the prefrontal cortex is more important in higher intellectual functions than other portions of the cortex have not been successful. Indeed, destruction of the posterior temporal lobe (Wernicke's area) and the angular gyrus region in the dominant hemisphere causes infinitely more harm to the intellect than does destruction of both prefrontal areas

Yet, the prefrontal areas do have some specific functions that are all their own One of these has to do with control of some types of behavior, especially choice of behavioral options for each social or physical situation. For this purpose, the prefrontal areas transmit signals into the limbic areas of the brain, as will be described in Chapter 56. Some of the other possible functions of the prefrontal lobes include

Prevention of Distractibility—Importance for Sequencing Thoughts. One of the outstanding characteristics of a person who has lost the prefrontal areas is the ease with which he or she can be *distracted* from a sequence of thoughts Likewise, in lower animals whose prefrontal areas have been removed, the ability to concentrate on psychological tests is almost completely lost. The human being without prefrontal areas is still capable of performing many intellectual tasks, such as answering short questions and performing simple arithmetic computations (such as $9 \times 6 = 54$), thus illustrating that the basic intellectual activities of the cerebral cortex are still intact without the prefrontal areas Yet if concerted *sequences* of cerebral functions are required of the person, he becomes completely disorganized. Therefore, the prefrontal areas seem to be important in keeping the mental functions directed toward goals.

Elaboration of Thought, Prognostication, and Performance of Higher Intellectual Functions by the Prefrontal Areas. Another function that has been ascribed to the prefrontal areas by psychologists and neurologists is *elaboration of thought*. This means simply an increase in depth and abstractness of the different thoughts. Psychological tests have shown that prefrontal lobectomized lower animals presented with successive bits of sensory information fail to store these bits even in tempory memory—probably because they are distracted so easily that they cannot hold thoughts long enough for storage to take place. This ability of the prefrontal areas to cause storage—even though it be temporary—of many bits of information simultaneously, and then to cause recall of this information bit by bit as it is needed for subsequent thoughts, could well explain the many functions of the brain that we associate with higher intelligence, such as the abilities to (1) prognosticate, (2) plan for the future, (3) delay action in response to incoming sensory signals so that the sensory information can be weighed until the best course of response is decided, (4) consider the consequences of motor actions even before these are performed, (5) solve complicated mathematical, legal, or philosophical problems, (6) correlate all avenues of information in diagnosing rare diseases, and (7) control one's activities in accord with moral laws.

Effects of Destruction of the Prefrontal Areas

Human beings without prefrontal areas ordinarily act precipitously in response to incoming sensory signals, such as reacting angrily to slight provocations but also forgetting the event equally as rapidly In addition, they are likely to lose many or most of their morals; they have little embarrassment in relation to excretory, sexual, and social activities; and they are prone to quickly changing moods of sweetness, hate, joy, sadness, exhilaration, and rage. And they lose both ambition and drive toward goals. In short, they are highly *distractible* persons with lack of ability to pursue long and complicated thoughts.

THOUGHTS, CONSCIOUSNESS, AND MEMORY

Our most difficult problem in discussing consciousness, thoughts, memory, and learning is that we do not know the neural mechanism of a thought. We know that destruction of large portions of the cerebral cortex does not prevent a person from having thoughts, but it usually does reduce the *degree* of awareness of the surroundings.

Each thought almost certainly involves simultaneous signals in many portions of the cerebral cortex, thalamus, limbic system, and reticular formation of the brain stem. Some crude thoughts probably depend almost entirely on lower centers;

the thought of pain is probably a good example, for electrical stimulation of the human cortex rarely elicits anything more than the mildest degrees of pain, whereas stimulation of certain areas of the hypothalamus and mesencephalon in animals appears to cause excruciating pain. On the other hand, a type of thought pattern that requires mainly the cerebral cortex is that involving vision, because loss of the visual cortex causes complete inability to perceive visual form or color.

Therefore, we might formulate a definition of a thought in terms of neural activity as follows: A thought probably results from the momentary "pattern" of stimulation of many different parts of the nervous system at the same time, probably involving most importantly the cerebral cortex, the thalamus, the limbic system, and the upper reticular formation of the brain stem. This is called the *holistic theory* of thoughts. The stimulated areas of the limbic system, thalamus, and reticular formation perhaps determine the general nature of the thought, giving it such qualities as pleasure, displeasure, pain, comfort, crude modalities of sensation, localization to gross areas of the body, and other general characteristics. On the other hand, the stimulated areas of the cerebral cortex probably determine the discrete characteristics of the thought such as specific localization of sensations on the body and of objects in the fields of vision, discrete patterns of sensation such as the rectangular pattern of a concrete block wall or the texture of a rug, and other individual characteristics that enter into the overall awareness of a particular instant.

And consciousness can perhaps be described as our continuing stream of awareness of either our surroundings or our sequential thoughts.

MEMORY AND TYPES OF MEMORY

If we accept the above approximation of what constitutes a thought, we can see immediately that the mechanism of memory must be equally as complex as the mechanism of a thought, for, to provide memory, the nervous system must recreate the same spatial and temporal pattern (the "holistic" pattern) of stimulation in the central nervous system at some future date. Though we cannot explain in detail what a memory is, we do know some of the basic psychological and neuronal processes that probably lead to the process of memory.

All of us know that all degrees of memory occur, some memories lasting a few seconds and others lasting hours, days, months, or years. Possibly all these types of memory are caused by the same mechanism operating to different degrees of fulfillment. Yet, it is also possible that different mechanisms of memory do exist. Indeed, most physiologists classify memory into from two to four different types. For the purpose of the present discussion, we will use the following classification:

1. *Sensory memory;*
2. *Primary memory;*
3. *Secondary memory.*

The basic characteristics of these types of memory are the following:

Sensory Memory. Sensory memory means the ability to retain sensory signals in the sensory areas of the brain for a very short interval of time following the actual sensory experience. Usually these signals remain available for analysis for several hundred milliseconds but are replaced by new sensory signals in less than one second. Nevertheless, during the short interval of time that the instantaneous sensory information remains in the brain it can continue to be used for further processing; most important, it can be "scanned" to pick out the important points. Thus, this is the initial stage of the memory process.

Primary Memory. Primary memory is the memory of facts, words, numbers, letters, or other information for a few seconds to a few minutes at a time. This is typified by a person's memory of the digits in a telephone number for a short period of time after he has looked up the number in the telephone directory. It is also typified by the ability of a person to look at a visual scene, then to turn the head away and still be able to recall for seconds or a minute or more many features of the scene.

One of the most important characteristics of primary memory is that the information in this memory store is instantaneously available so that the person does not have to search through his or her mind for it as one does for information that has been put away in the secondary memory stores. However, still another feature of primary memory is that when new bits of information are put into the primary store, old information is displaced. Thus, if a person looks up a second telephone number, the first is usually lost. Also, if one sees a rapid succession of visual scenes, it is usually the last of these that remains most prominently in the primary memory store.

Secondary Memory. Secondary memory is the storage in the brain of information that can be recalled at some later time—hours, days, months, or years later. This type of memory has been called *long-term memory, fixed memory, permanent memory,* and several other names. One of its characteristics is that, except when the memory is very deeply engrained, one must "search" through the memory stores for seconds to minutes before it is possible to recall the memory.

"Short-term" and "Long-term" Memory. Memory is also frequently divided into either "short-term" or "long-term." However, use of these terms is extremely loose. Psychologists frequently classify short-term memory to be the same as primary memory and long-term memory the same as secondary memory. However, many physiologists and clinicians include in short-term memory the early stages of secondary memory lasting for as long as several days to a week or more, and reserve the term long-term memory for memories that can be re-

called weeks, months, or years after the initial experience

PHYSIOLOGICAL BASIS OF MEMORY

Despite the many advances in neurophysiology during the past half century, we still cannot explain what is perhaps the most important function of the brain: its capability for memory. Yet, physiological experiments are beginning to generate conceptual theories of the means by which memory could occur. Some of these are discussed in the following few sections.

Possible Mechanisms for Primary Memory. Primary memory requires a neuronal mechanism that can hold specific information signals for a few seconds to at most a minute or more. Several such mechanisms are the following:

Reverberating Circuit Theory of Primary Memory. When an appropriate tetanizing electrical stimulus is applied directly to the surface of the cerebral cortex and then is removed after a second or more, the local area excited by this stimulus often continues to emit rhythmic action potentials for short periods of time. This effect is believed to result from local reverberating circuits, the signals passing through a multistage circuit of neurons in the local area of the cortex itself or perhaps even back and forth between the cortex and the thalamus or other subcortical areas.

It is postulated that sensory signals reaching tha cerebral cortex can set up similar reverberating oscillations and that these could be the basis for primary memory. Then, as the reverberating circuit fatigues, or as new signals interfere with tha reverberations, the primary memory fades away.

One of the principal observations in support of this theory is that any factor that suddenly blocks brain function, such as a blow on the head that causes temporary coma or such as anesthesia, erases the primary memory. The memory cannot be recalled when the disturbance is over unless a portion of this memory had already been placed into the secondary memory store, as will be discussed in subsequent sections

Post-Tetanic Potentiation Theory of Primary Memory. In most parts of the nervous system, including even the anterior motor neurons of the spinal cord, tetanic stimulation of a synapse for a few seconds causes a short period of synaptic fatigue for the next few seconds but then increased excitability of the synapse for another few seconds to a few hours. If during this time of increased excitability the synapse is stimulated again, the neuron responds much more vigorously than normally, a phenomenon called *post-tetanic potentiation*. This is obviously a type of memory that depends on change in the excitability of the synapse, and it could be the basis for primary memory. Recent experiments have shown that post-tetanic potentiation is caused by excessive accumulation of calcium ions in the presynaptic terminals, these ions in turn increasing the release of transmitter substance by the terminals, as was explained in Chapter 46.

Presynaptic Facilitation and Inhibition as Possible Causes of Primary Memory. Many presynaptic terminals of the central nervous system lie on the surfaces of other presynaptic terminals or even on terminal nerve fibrils When such terminals are stimulated, they can cause either inhibition or facilitation of the secondary terminals depending on the type of transmitter released. Furthermore, this inhibition or facilitation can last at times for many minutes and perhaps even for an hour or more. Such events are a type of memory, though it still is not known whether this is one of the mechanisms of primary memory.

Mechanism of Secondary Memory—Enhancement of Synaptic Transmission Facility

Secondary memory is the ability of the nervous system to recall thoughts long after initial elicitation of the thoughts is over. We know that secondary memory does not depend on continued activity of the nervous system, because the brain can be totally inactivated by cooling, by general anesthesia, by hypoxia, by ischemia, or by any other method and yet secondary memories that have been previously stored are still retained when the brain becomes active once again Therefore secondary memory must result from some actual alterations of the synapses, either physical or chemical.

Many different theories have been offered to explain the synaptic changes that cause long-term memory. Among the most important of these are:

1. Anatomical Changes in the Synapses. Cajal, almost a century ago, discovered that the number of terminal fibrils ending on neuronal cells and dendrites in the cerebral cortex increases with age. Conversely, physiologists have shown that inactivity of regions of the cortex causes thinning of the cortex: for instance, thinning of the primary visual cortex in animals that have lost their eyesight. Also, intense activity of a particular part of the cortex can cause excessive thickening of the cortical shell in that area alone. This has been demonstrated especially in the visual cortex of animals subjected to repeated visual experiences. Finally, some neuroanatomists have observed electron micrographic changes in presynaptic terminals that have been subjected to intense and prolonged activity.

All these observations have led to a widely held belief that fixation of memories in the brain results from anatomical changes in the synapses themselves: perhaps changes in numbers of presynaptic

terminals, perhaps in sizes of the terminals, or perhaps in the sizes and conductivities of the dendrites. Such anatomical changes could cause permanent or semipermanent increase in the degree of facilitation of specific neuronal circuits, thus allowing signals to pass through the circuits with more ease the more often the memory trace is used. This obviously would explain the tendency for memories to become more and more deeply fixed in the nervous system the more often they are recalled or the more often the person repeats the sensory experience that leads to the memory trace.

Another variant of the anatomical theory is based on the experimental observation that in early life an excess of synaptic connections is made, and those snyapses that are active continue to remain functional whereas the others eventually degenerate. Thus, the pattern of synaptic transmission could be established by destroying unused synapses rather than by enhancing the used ones. This theory fits the finding that unused synapses and even unused neurons often disappear completely.

2. Physical or Chemical Changes in the Presynaptic Terminal or the Postsynaptic Membrane. Recent studies by Kandel and others in the large snail *Aplysia* have uncovered several mechanisms of memory that result from either physical or chemical changes in the presynaptic terminal or possibly even in the entire presynaptic neuron. One such mechanism, illustrated in Figure 54–7, functions in the following way: In this figure there are two separate presynaptic terminals, one of which is from a primary input sensory neuron and terminates on the surface of a secondary neuron that is to be stimulated; this is called the *sensory terminal*. The other presynaptic terminal lies on the surface of the sensory terminal and is called the *facilitator terminal*. When the sensory terminal is stimulated repeatedly but without stimulating the facilitator terminal, signal transmission at first is very great, but this becomes less and less intense with repeated stimulation until transmission almost ceases. This phenomenon is called *habituation*. It is a type of memory that causes the neuronal circuit to lose its response to repeated events that are insignificant.

On the other hand, if a noxious stimulus excites the facilitator terminal at the same time that the sensory terminal is stimulated, then, instead of the transmitted signal becoming progressively weaker, the ease of transmission becomes much stronger and will remain strong for hours, days, or perhaps weeks even without further stimulation of the facilitator terminal. Thus, the noxious stimulus causes the memory pathway to become facilitated for days or weeks thereafter. It is especially interesting that only a few repeated action potentials are required to cause either habituation or facilitation. However, once habituation has occurred, the synapses can become facilitated very rapidly with only a few noxious stimuli.

At the molecular level, the habituation effect in the sensory terminal results from progressive closure of calcium channels of the terminal membrane, though the cause of this is not fully known. As a result, much smaller than normal amounts of calcium diffuse into this terminal when action potentials occur, and much less transmitter is therefore released because calcium entry is the stimulus for transmitter release (as was discussed in Chapter 46).

In the case of facilitation, the molecular mechanism is believed to be the following:

(1) Stimulation of the facilitator neuron at the same time that the sensory neuron is stimulated causes serotonin release at the facilitator synapse on the sensory presynaptic terminal.

(2) The serotonin acts on *serotonin receptors* in the sensory terminal membrane, and these activate the enzyme *adenylate cyclase* inside the membrane. This causes the formation of *cyclic AMP* inside the sensory presynaptic terminal.

(3) The cyclic AMP activates a *protein kinase* that is believed to cause phosphorylation of the channel surfaces of the potassium channels in the sensory terminal membrane. This blocks these channels for potassium conductance. This blockage of the potassium channels can last for days and maybe even weeks or longer.

(4) Lack of potassium conductance causes a greatly prolonged action potential in the presynaptic terminal because the flow of potassium ions out of the terminal is the usual mechanism for recovery from the action potential.

(5) The prolonged action potential causes prolonged activation of the calcium pores, allowing tremendous quantities of calcium ions to enter the sensory terminal. These calcium ions then cause greatly increased transmitter release, thereby greatly facilitating synaptic transmission.

Thus, in a very indirect way the associative effect of stimulating the facilitator neuron at the same time that the sensory neuron is stimulated causes a prolonged change in the sensory terminal that produces a long-term memory trace.

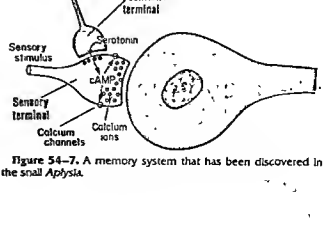


Figure 54–7. A memory system that has been discovered in the snail *Aplysia*.

Recent studies by Byrne in the snail Aplysia have suggested still another mechanism of cellular memory; his studies have shown that stimuli from two separate sources acting on a single sensory neuron can, under appropriate conditions, cause long-term changes in the membrane properties of the entire sensory neuron, and these changes can increase the excitability of the neuron for many minutes thereafter. Thus, this is another possible mechanism of memory, either primary memory or secondary memory

Summary. The theory that at present seems most likely for explaining secondary memory is that some actual anatomical, physical, or chemical change occurs in the presynaptic terminals or perhaps in whole neurons that permanently facilitate the transmission of impulses at the synapses. If all the synapses are thus facilitated in a thought circuit, this circuit can be re-excited by any one of many diverse signals at later dates, thereby causing memory. The overall facilitated circuit is called a *memory engram* or a *memory trace.*

Consolidation of Memory

For a primary memory to be converted into a secondary memory that can be recalled days later, it must become "consolidated"—that is, the synapses must become permanently facilitated This process requires 5 to 10 minutes for minimal consolidation and an hour or more for maximal consolidation. For instance, if a strong sensory impression is made on the brain but is then followed within a minute or so by an electrically induced brain convulsion, the sensory experience will not be remembered at all Likewise, brain concussion, sudden application of deep general anesthesia, and other effects that temporarily block the dynamic function of the brain can prevent consolidation

However, if the same sensory stimulus is impressed on the brain and the strong electrical shock is delayed for more than 5 to 10 minutes, at least part of the memory trace will have become established. If the shock is delayed for an hour, the memory will have become fully consolidated

The process of consolidation and the time required for consolidation can probably be explained by the phenomenon of *rehearsal* of the primary memory as follows:

Role of Rehearsal in Transference of Primary Memory into Secondary Memory. Psychological studies have shown that rehearsal of the same information again and again accelerates and potentiates the degree of transfer of primary memory into secondary memory, and therefore also accelerates and potentiates the process of consolidation. The brain has a natural tendency to rehearse newfound information, and especially to rehearse newfound information that catches the mind's attention. Therefore, over a period of time the important features of sensory experiences become progressively more and more fixed in the secondary memory stores. This explains why person can remember small amounts of inform tion studied in depth far better than large amoun of information studied only superficially. And also explains why a person who is wide-awake w consolidate memories far better than will a pers who is in a state of mental fatigue.

Codifying of Memories During the Proce of Consolidation. One of the most importa features of the process of consolidation is th memories to be placed permanently into the se ondary memory storehouse are first codified in different classes of information. During this pr ess similar information is recalled from the se ondary storage bins and is used to help process t new information The new and old are compar for similarities and for differences, and part of t storage process is to store the information ab these similarities and differences rather than si ply to store the information unprocessed. Th during the process of consolidation, the new me ories are not stored randomly in the brain b instead are stored in direct association with oth memories of the same type. This is obviously n essary if one is to be able to "search" the memory store at a later date to find the required information.

Role of Specific Parts of the Brain in the Memory Process

Role of the Hippocampus for Rehearsal, Codification, and Consolidation of Memories—Anterograde Amnesia Following Hippocampal Lesions. The hippocampus is the most medial portion of the temporal lobe cortex where it folds underneath the brain and then upward into the lower surface of the lateral ventricle. The two hippocampi have been removed for the treatment of epilepsy in a number of patients. This procedure hardly affects the person's memory for information stored in the brain prior to removal of the hippocampi. However, after removal, these persons have very little capability for transferring *verbal and symbolic types* of primary memories into secondary memories. That is, they do not have the ability to separate out the important information; to codify it, to rehearse it, and to consolidate it in the long-term memory store. Therefore, these persons are unable to establish new long-term memories of those types of information that are the basis of intelligence. This is called *anterograde amnesia.*

Lesions in other parts of the temporal lobes besides the hippocampi are also frequently associated with reduced ability to store new memories. This probably results from two factors: (1) the possible association of the other parts of the temporal lobes with the hippocampi and therefore failure of the usual consolidation process for mem-

ories, and (2) the fact that Wernicke's area, which is the major locus of intellectual operations of the brain, is located in the temporal lobe; the reason that lesions affecting Wernicke's area might diminish memory storage is probably that consolidation of memories requires analysis of the memory so that it can be stored in association with other memories of like kind.

Retrograde Amnesia. *Retrograde amnesia* means inability to recall memories from the past—that is, from the long-term memory storage bins—even though the memories are known to be still there. When retrograde amnesia occurs, the degree of amnesia for recent events is likely to be much greater than for events of the distant past. The reason for this difference is probably that the distant memories have been rehearsed so many times that elements of these memories are stored in widespread areas of the brain

In some persons who have hippocampal lesions, some degree of retrograde amnesia usually occurs along with the anterograde amnesia just discussed, which suggests that these two types of amnesia are at least partially related and that hippocampal lesions can cause both. However, it has also been claimed that damage in some thalamic areas can lead specifically to retrograde amnesia without causing significant anterograde amnesia. This perhaps can be explained by the possibility that the thalamus might normally play a role in helping the person "search" the memory storehouses and thus be able to "read out" the memories. That is, the memory process not only requires the storing of memories but also the ability to search and find the memory at a later date. The possible function of the thalamus in this process will be discussed in the following chapter.

It should be noted, however, that persons with either temporal lobe or hippocampal lesions usually do not have difficulty in learning physical skills that do not involve verbalization or symbolic types of intelligence. For instance, these persons can still learn hand skills such as those required in carpentry.

ANALYTICAL OPERATIONS OF THE BRAIN

Thus far, we have considered the approximate nature of thoughts and possible mechanisms by which memory and learning can occur Now we need to consider the mechanisms by which the brain performs complex intellectual operations, such as the analysis of sensory information and the establishment of abstract thoughts. About these mechanisms we know almost nothing, but experiments along these lines have established the following important facts: First, the brain can focus its attention on specific types of information (which will be discussed in the following chapter). Second, separate qualities of each set of information signals are split away from the central signal and are transmitted to multiple areas of the brain. Third, the brain compares new information with old information in its memory loci. And, fourth, the brain determines patterns of stimulation.

Analysis of Information by Splitting Its Qualities. When sensory information enters the nervous system, one of the first steps in its analysis is to transmit the signal to separate parts of the brain that are selectively adapted for detecting specific qualities. For instance, if the hand is placed on a hot stove, (a) pain information is sent through the spinoreticular and spinothalamic tracts into the reticular formation of the brain stem and into certain thalamic and hypothalamic nuclei, (b) tactile information is sent through the dorsal column system to the somatic sensory cortex, giving the cortex a detailed description of the part of the hand that is touching the hot stove, and (c) position sense information from the muscles and joints is sent to the cerebellum, the reticular formation, and to higher centers of the brain to give momentary information of the position of the hand Thus, the different qualities of the overall information are dissected and transmitted to different parts of the brain On the basis of analyses in all these parts of the brain, motor responses to the incoming sensory information are formulated

One can see that different centers of the brain react to specific qualities of information—the reticular formation and certain regions of the thalamus and hypothalamus to pain, regions of the mesencephalon and hypothalamus to the affective nature of the sensation (that is, whether it is pleasant or unpleasant), the somatic sensory cortex to the localization of sensation, other areas to kinesthetic activities, and others to visual, auditory, gustatory, olfactory, and vestibular information.

Analysis of New Information by Comparison with Memories. We all know that new sensory experiences are immediately compared with previous experiences of the same or similar types. For instance, this is the way we recognize a person whom we know Yet, it still remains a mystery how we can make these comparisons One theory suggests that if the new pattern of stimulation matches a memory engram, some interaction between the two gives the person a sense of recognition.

Analysis of Patterns. In analyzing information, the brain depends to a great extent on "patterns" of stimulation. For instance, a square is detected as a square regardless of its position or angle of rotation in the visual field Likewise, a series of parallel bars is detected as parallel bars regardless of their orientation, or a fly is detected as a fly whether it is seen in the peripheral or central field of vision. We can extend the same logic to the somatic sensory areas, for a person can detect a cube simply by feeling, whether it be in an upright, horizontal, or angulated position. Also, it can be detected even by one's feet even though one's feet may have never felt a cube before.

Processing of Information so That Patterns Can Be Determined. Beautiful examples of the ability of the brain to process information for determination of patterns have been discovered in relation to the visual system. In Chapter 60 we shall point out that the retina itself processes visual information to a great extent even before it is transmitted into the brain. For instance, the neuronal circuits in the retina are organized to detect lines or boundaries between light and dark areas. That is, the visual neural mechanisms bring out strongly the

contrast boundaries of the scene but de-emphasize the flat areas. This explains why a few lines drawn on a paper can give one the impression of seeing the image of a person Certainly the lines do not represent the actual picture of the person, but they do give the visual cortex the same pattern of contrasts that one's own visual system would give

One can see, therefore, that the visual cortex utilizes transformed types of information to point up the most striking characteristics of the visual scene while neglecting unimportant information Then, the patterns of the visual scene can be extracted from this "preselected" information

FUNCTION OF THE BRAIN IN COMMUNICATION

One of the most important differences between the human being and lower animals is the facility with which human beings can communicate with one another Furthermore, because neurological tests zan easily assess the ability of a person to communicate with others, we know perhaps moco about the sensory and motor systems related to communication than about any other segment of cortical function. Therefore, we will review rapidly the function of the cortex in communication, and from this ono can see immediately how the principles of sensory analysis and motor conteol apply to this art

There are two aspecto to communication: first, the *sensory aspect*, involving the ears and eyea, and, second, the *motor aspect*, involving vocalization and its contral

Sensory Aspects of Communication. We noted earlier in the chapter that destruction of portiona of the *auditory* and *visual association areas* of the cortex can rasult in inability to understand the spoken word or the written word. These effects are called, respectively, *auditory receptive aphasia* and *visual receptive aphasia* or, more commonly, *word deafness* and *word blindness* (also called *dyslexia)*

Wernicke's Aphasia (Sensory Aphasia). Some persons are perfectly capable of understanding either the spoken word or the written word but are *unable to* interpret the thought that is expressed. This results most frequently when *Wernicke's area* in the *posterior portion of the dominant hemisphere superior temporal gyrus* is damaged or destroyed Therefore, this type of aphasia is generally called *Wernicke's aphasia* also, it is frequently called simply "sensory aphasia."

When the lesion in Wernicke's area is widespread and extends (a) backward into the angular gyrus region, (b) inferiorly into the lower areas of the temporal lobe, and (c) superiorly into the superior border of the sylvian fissure, the person is likely to be almost totally demented and therefore is said to have *global aphasia.*

Motor Aspects of Communication. The process of speech involves two principal stagea of mentation: (1) formation in the mind of thoughts to be expressed and choice of words to be used, then (2) motor control of vocalization and the actual act of vocalization itself The formation of thoughts and choice of words is the function of the sensory areas of the brain. Again, it is Wernicke's area in the posterior part of the superior temporal gyrus that is most important for this ability. Therefore, persons with either Wernicke's aphasia or global aphasia are unable to formulate the thoughts that are to be communicated Or, if the lesion is less severe, the perso may be able to formulate thoughts but yet be unable t put together the appeopriate words to express th thought. Often, a person may be very fluent in word but the words may present a jumbled mass of confusion

Motor Aphasia. Often a person is perfectly capabl of deciding what he wishes to say, and he is capable o vocalizing, but he simply cannot make his vocal systen emit words instead of noises. This effect, called *moto aphasia*, results feom damage to *Broca's speech are* which lies in the *premotor* facial region of the cortex about 95 per cent of the time in the left hemisphere, a illustrated in Figures 54–5 and 54–8. Therefore, w assume that the *skilled motor patterns* for control of th larynx, lips, mouth, respiratory system, and other ac cessory muscles of articulation are all controlled in th area

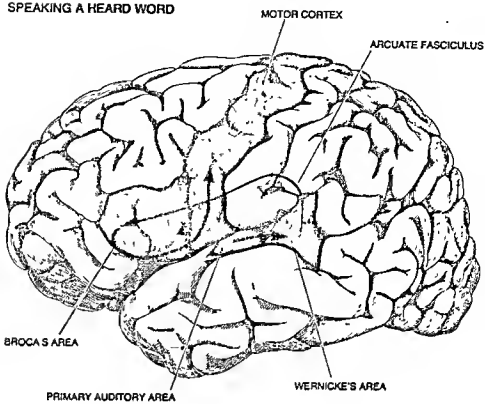
Articulation. Finally, we have the act of articulation itself, which means the muscular movements of th mouth, tongue, larynx, and so forth, that are responsibl for the actual emission of sound The *facial and laryngea regions of the motor cortex* activate these muscles, an the *cerebellum, basal ganglia*, and *sensory cortex* all hel contral the muscle contractions by feedback mechanism described in Chapter 53 Destruction of these region can couse either total or partial inability to spea distinctly

Summary. Figure 54–8 illustrates two principal pati ways for communication The upper half of tha figu shows the pathway involved in hearing and speaking word This sequence is the following: (1) reception i the primary auditory area of the sound signals th encode the word, (2) interpretation of the word in We nicke's area; (3) determination also in Wernicke's are that the word is to be spoken; (4) transmission of signal feom Wernicke's area to Broca's area via the *arcua fasciculus*, (5) activation of the skilled motor progran in Broca's area for control of word formation; and (6 transmission of appeopriate signals into the motor corte to conteol the speech muscles

The lower figure illustrates the comparable steps i reading a word and then speaking it. The initial recep tive area for the word is in the primary visual are rather than in the primary auditory area Then the wor information passes through early stages of interpreta tian in the *angular gyrus region* and finally reaches i full level of recognition in Wernicke's area. From her the sequence is the same as for speaking a heard wor

FUNCTION OF THE CORPUS CALLOSUM AND ANTERIOR COMMISSURE TO TRANSFER THOUGHTS, MEMORIES, AND OTHER INFORMATION TO THE OPPOSITE HEMISPHERE

Fibers in the *corpus callosum* connect the respectiv cortical areas of the two hemispheres with each othe except for the anterior portions of the temporal lobe these temporal areas, including especially the *amygdal* are interconnected by fibers that pass through the *an terior commussure.* Because of the tremendous numbe of fibers in the corpus callosum, it was assumed fro the beginning that this massive structure must hav



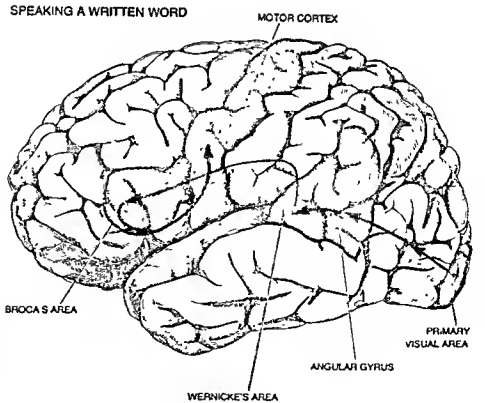


Figure 54–8. Brain pathways for *(top)* perception of the heard word and then speaking the same word, and *(bottom)* perception of the written word and then speaking the same word. (From Geschwind, *Sci. Am.*, *241*:180, 1979. © 1979 by Scientific American, Inc. All rights reserved.)

some important function to correlate activities of the two cerebral hemispheres However, after cutting the corpus callosum in experimental animals, it was difficult to discern changes in brain function Therefore, for a long time the function of the corpus callosum was a mystery

Yet, properly designed psychological experiments have now demonstrated the extremely important functions of the corpus callosum and anterior commissure. These can be explained best by recounting one of the experiments A monkey is first prepared by cutting the corpus callosum and splitting the optic chiasm longitudinally Then it is taught to recognize different types of objects with its right eye while its left eye is covered. Next, the right eye is covered and the monkey is tested to determine whether or not its left eye can recognize the same object. The answer to this is that the left eye *cannot* recognize the object Yet, on repeating the same experiment in another monkey with the optic chiasm split but the corpus callosum intact, it is found invariably that recognition in one hemisphere of the brain creates recognition also in the opposite hemisphere

Thus, one of the functions of the corpus callosum and the anterior commissure is to make information stored in the cortex of one hemisphere available to the corresponding cortical area of the opposite hemisphere Three important examples of such cooperation between the two hemispheres are·

1. Cutting of the corpus callosum blocks transfer of information from Wernicke's area of the dominant hemisphere to the motor cortex on the opposite side of the brain Therefore, the intellectual functions of the brain, located primarily in the dominant hemisphere, lose their control over the right motor cortex and therefore also of the voluntary motor functions of the left hand and arm even though the usual subconscious movements of the left hand and arm are completely normal

2 Cutting of the corpus callosum prevents transfer of somatic and visual information from the right hemisphere into Wernicke's area of the dominant hemisphere Therefore, somatic and visual information from the left side of the body frequently fails to resch this general interpretative area of the brain and, therefore, cannot be used for decision making

3 Finally, a teen-aged boy whose corpus callosum had been completely sectioned but whose anterior commissure was not sectioned recently was found to have two entirely separate conscious portions to his brain. Only the left half of his brain could understand the spoken word, because it was the dominant hemisphere. On the other hand, the right side of the brain could understand the written word and could elicit a motor response to it without the left side of the brain ever knowing why the response was performed Yet, the effect was somewhat different when an emotional response was evoked in the right side of the brain in this case a subconscious emotional response occurred in the left side of the brain as well This undoubtedly occurred because the temporal lobe areas and other basal brain areas of the two sides of the brain were still communicating with each other through the anterior commissure that was not sectioned. For instance, when the command "kiss" was written for the right half of his brain to see, the boy immediately and with full emotion, said "No way!" This response obviously required function of Wernicke's area and the motor areas for speech in the left hemisphere But, when questioned why he said this, the boy could not explain. Thus, the two halves of the brain have independent capabilities for consciousness, memory storage, communication, and control of motor activities The corpus callosum is required for the two sides to operate cooperatively, and the anterior commissure plays an important role in unifying the emotional responses of the two brain sides.

REFERENCES

Bannister R. (rev) Brain's Clinical Neurology New York, Oxford University Press, 1978

Bekhtereva, N P The Neurophysiological Aspects of Human Mental Activity New York, Oxford University Press, 1978

Benson, D F Aphasia, Alexia and Agraphia. New York, Churchill Livingstone, 1979

Bindman, L The Neurophysiology of the Cerebral Cortex. Austin, University of Texas Press, 1981

Buchtel, H A The Conceptual Nervous System New York, Pergamon Press, 1982.

Buser, P Higher functions of the nervous system *Annu. Rev Physiol.* *38* 217, 1976

Damasio, A R, and Geschwind, N The neural basis of language. *Annu. Rev. Neurosci*, *7* 127 1984

Damloff, R., *et al* The Physiological Bases of Verbal Communication. Englewood Cliffs *NJ*, Prentice Hall, 1980

De Silva, F H L, and Arnolds, D E.A.T., Physiology of the hippocampus and related structures *Annu Rev Physiol*, *40*·185, 1978

Dimond, S J Neuropsychology· A Textbook of Systems and Psychological Functions of the Human Brain Boston, Butterworths, 1979

Edelman, G M, and Mountcastle, V B The Mindful Brain. Cortical Organization and the Group-Selective Theory of Higher Brain Function Cambridge, Mass MIT Press, 1978

Geschwind, N Specializations of the human brain *Sci. Am*, *241*(3)·180, 1979

Hixon, T J, *et al* (eds) Introduction to Communicative Disorders. Englewood Cliffs, *NJ*, Prentice-Hall, 1980

Hubel, D H The brain *Sci Am*, *241*(3) 44, 1979

Hyvarinen J Posterior parietal lobe of the primate brain. *Physiol Rev*, *62* 1060, 1982

Kandel, E R Neuronal plasticity and the modification of behavior *In* Brookhart, J M, and Mountcastle, V B (eds)· Handbook of Physiology Sec 1, Vol I, Baltimore Williams & Wilkins, 1977, p 1137

Kandel E R A Cell Biological Approach to Learning Bethesda, Md., Society for Neuroscience 1978

Klopf A H The Hedonistic Neuron A Theory of Memory, Learning and Intelligence Washington, D C, Hemisphere Publishers, 1982.

Laduron, P M Presynaptic heteroreceptors in regulation of neuronal transmission *Biochem Pharmacol*, *34* 467, 1985

McCloskey, D I. Corollary discharges Motor commands and perception. *In* Brooks, V B (ed)· Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 1415

McNeil, M R (ed)· The Dysarthrias Physiology, Acoustics, and Perception Management San Diego, College-Hill Press, 1984

Mitzdorf, U Current source-density method and application in cat cerebral cortex Investigation of evoked potentials and EEG phenomena *Physiol Rev* *65* 37, 1985

Moskowitz, B A The acquisition of language *Sci Am*, *239*(5) 92, 1978.

Quinn, W G, and Greenspan, R J Learning and courtship in Drosophila. Two stories with mutants. *Annu Rev Neurosci*, *7* 67, 1984

Rosenbek, J C (ed)· Apraxia of Speech Physiology, Acoustics, Linguistics, Management San Diego, College-Hill Press, 1984

Sperry, R. W., Changing concepts of consciousness and free will *Perspect. Biol. Med.*, *20* 9, 1976

Thompson, R F, *et al* Cellular processes of learning and memory in the mammalian CNS *Annu. Rev Neurosci*, *6* 447, 1983

Trevarthen, C Hemispheric specialization *In* Darian-Smith, I. (ed.)· Handbook of Physiology Sec 1, Vol III Bethesda, American Physiological Society, 1984, p 1129

Truman, J W · Cell death in invertebrate nervous systems. *Annu. Rev Neurosci*, *7* 171, 1984

Uttal, W R. The Physiology of Mind. New York, Halsted Press, 1978.

Vellutino, F R. Dyslexia Theory and Research Cambridge, Mass., MIT Press, 1979

Walters E T, and Byrne, J H., Associative conditioning of single sensory neurons suggests a cellular mechanism for learning Science *219* 405, 1983

Wong, R K, *et al*. Local circuit interactions in synchronization of cortical neurones *J Exp Biol*, *112* 169, 1984

55

Activation of the Brain—the Reticular Activating System and Wakefulness; Attention and Searching of Memory; Brain Waves; Sleep; Epilepsy

One of the remaining great mysteries of the brain is how the brain itself controls itself. For instance, what sets the overall level of activity? Also, why do we go to sleep or wake up? And perhaps even more mysterious is the mechanism by which we can focus our attention on certain functions of the brain while ignoring others.

Though the answers to these questions are mainly unknown, glimmers of information are beginning to appear that will at least let us construct plausible, even though questionable, theories about activation of the brain, wakefulness, sleep, and so forth.

THE RETICULAR ACTIVATING SYSTEM AND ITS ROLE IN WAKEFULNESS

Diffuse electrical stimulation in the *mesencephalic and pontile portions of the reticular formation*—an area discussed in Chapter 52 in relation to the motor functions of the nervous system—causes immediate and marked activation of the cerebral cortex and will even cause a sleeping animal to awaken instantaneously. This area is represented in Figure 55–1 by the bold arrow directed upward through the brain stem. Extending upward into the cerebrum from the mesencephalic reticular formation are multiple diffuse pathways that terminate in almost all areas of both the diencephalon and the cerebrum. This entire system is called the *reticular activating system.*

When the reticular activating system was first discovered almost 40 years ago, it was believed that this entire system was represented by a diffuse matrix of neurons that operated mainly as a single unit to provide generalized and uniform activation of all regions of the diencephalon and cerebrum at the same time. We now know, however, that certain regions of the reticular formation as well as portions of the upward radiating distribution system have discrete anatomical organizations, discrete distribution pathways, and different effects on different parts of the brain—at some times causing increased activity and at other times decreased activity—and even utilize different transmitter substances to perform these different functions. We shall discuss these more specific systems later, but for now let us address the more general functions.

Two Different Pathways into the Cerebrum from the Reticular Formation. Signals are transmitted from the mesencephalic portion of the

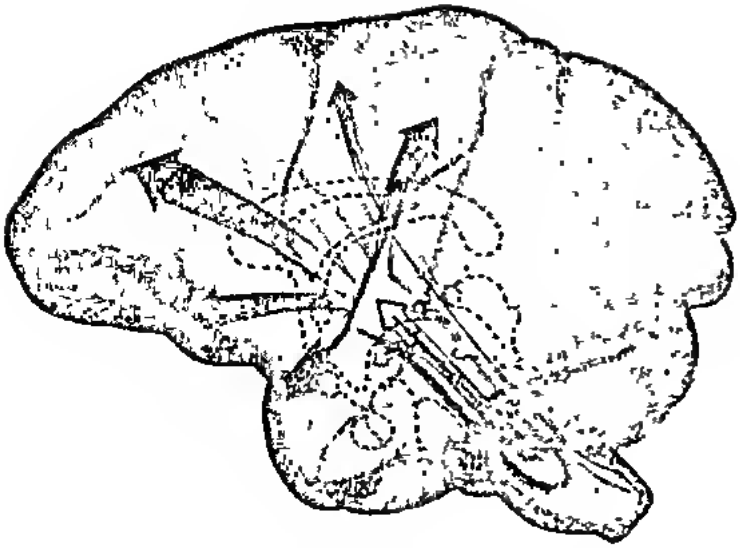


Figure 55–1. The reticular activating system schematically projected on a monkey brain. (From Lindsley: Reticular Formation of the Brain. Little, Brown and Co.)

reticular activating system to the cortex mainly through two different general pathways. One pathway passes upward to the intralaminar, midline, and reticular nuclei of the thalamus and thence through relay pathways to essentially all parts of the cerebral cortex and basal ganglia.

The second pathway, probably a less important pathway, passes through the subthalamus, hypothalamus, and adjacent areas.

Function of the Mesencephalic Portion of the Reticular Activating System. To understand the function of the mesencephalic portion of the reticular activating system it is instructive to consider two different experiments performed by cutting the brain stem at two different levels, as follows.

(1) Cutting the brain stem at the upper end of the mesencephalon causes the cerebrum to become inactive—that is, to go into coma.

(2) Cutting the brain stem in the midpontile region causes the cerebrum to become active and to remain active indefinitely as if it remains continuously awake.

Therefore, the reticular formation of the mesencephalon and upper pons seems to provide intrinsic activation of the brain. On the other hand, areas in the brain stem below the midlevel of the pons can inhibit this activating system and cause sleep.

Function of the Thalamic Portion of the Activating System. Electrical stimulation in different areas of the thalamic portion of the activating system activates specific regions of the cerebral cortex more than others. This is distinctly different from stimulation in the mesencephalic portion, which activates large areas of the brain at the same time. Therefore, it is postulated that selective stimulation of portions of the thalamus by the internal signals of the brain might be the cause of specific activation of certain areas of the cerebral cortex in distinction to other areas.

Stimulation of the Reticular Activating System by Nerve Signals from Elsewhere

The reticular activating system itself is subject to stimulation and therefore subject to increased levels of activation. Also, it is subject to inhibition which can lead to sleep, as we shall discuss in detail later in the chapter.

Two basic types of stimuli are especially likely to increase the activity of the activating system. These are

(1) Sensory stimuli from almost any part of the body.

(2) Retrograde stimuli from the cerebrum, feeding mainly into the mesencephalic portion of the reticular formation.

Almost any sensory signal entering the nervous system will cause at least some degree of activation of the reticular activating system. However, some signals are much more stimulatory than others, especially pain and proprioceptive somatic impulses, both of which are likely to require some immediate action by the brain. It will be recalled that the reticular formation receives tremendous numbers of signals either directly or through collaterals from the *spinoreticular tracts*, the *spinothalamic tracts*, and the *spinotectal tracts*. When an animal is asleep and appropriate sensory signals suddenly enter the reticular activating system, this can cause the animal to awaken immediately. This is called the *arousal reaction*.

Stimulation of the reticular activating system by the cerebrum, especially by the cerebral cortex, is mediated through direct fiber pathways into the reticular formation from almost all parts of the cerebrum, including (1) the somatic sensory cortex, (2) the motor cortex, (3) the frontal cortex, (4) the basal ganglia, (5) the hippocampus and other limbic structures, and (6) the hypothalamus. An especially large number of nerve fibers pass from the motor regions of the cerebral cortex to the

reticular formation; therefore, motor activity in particular is associated with a high degree of reticular activation, which partially explains the importance of moving around when one wishes to remain awake.

Wakefulness—Its Relationship to Activation of the Brain

We shall see later that the brain is exceedingly active in the course of generalized epileptic seizures, and yet the person is not awake. Likewise, the brain remains active to varying degrees during sleep and at some stages of sleep is perhaps equally as active as during wakefulness.

On the other hand, wakefulness does not occur in the absence of a reasonable degree of activation of the brain.

Therefore, wakefulness can be defined as activity in the brain directed into appropriate channels to give the person a sense of conscious awareness.

Coma

Coma is the opposite of brain activation. It can result from any factor that diminishes or stops activity in the mesencephalic portion of the reticular activating system. This can include such factors as:

(1) Brain tumors pressing on the brain stem, such as tumors of the pineal gland or the pituitary gland.

(2) Vascular lesions that interrupt the blood supply to the midbrain.

(3) Prolonged hypoxia that causes death of brain neurons.

(4) Infectious processes in the brain stem, such as the encephalitis that causes sleeping sickness.

(5) Poisons that destroy brain tissue.

Coma is distinct from sleep in that a person cannot be aroused from coma. In some coma patients all parts of the brain are inactivated, not just the reticular activating system. In this case all electrical activity of the brain ceases—that is, the brain waves are said to be "flat." This is the condition called *brain death,* and the person can then remain alive only by being sustained on artificial respiration, administration of nutrition by stomach tube or intravenously, and use of various supportive drugs and fluids to maintain appropriate blood circulation.

SPECIFIC NEURONAL SYSTEMS ASSOCIATED WITH THE RETICULAR FORMATION

We noted earlier that as more has been learned about the reticular formation, specific anatomically distinct neuronal systems have been discovered. Four of these have special importance in controlling specific activities of the diencephalon

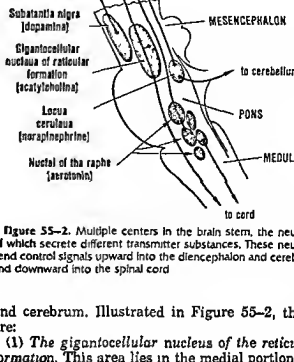


Figure 55–2. Multiple centers in the brain stem, the neurons of which secrete different transmitter substances. These neurons send control signals upward into the diencephalon and cerebrum and downward into the spinal cord

and cerebrum. Illustrated in Figure 55–2, these are:

(1) *The gigantocellular nucleus of the reticular formation.* This area lies in the medial portions of the reticular formation in the mesencephalon and upper pons. The neurons of this nucleus are very large and their fibers divide immediately into two branches, one passing upward into the higher levels of the brain and the other passing downward through the reticulospinal tracts into the spinal cord. This is a principal activator portion of the reticular activating system.

The neurons of this nucleus release *acetylcholine* at their terminals, and this normally functions as an excitatory transmitter.

(2) *The substantia nigra.* This nucleus was discussed in Chapter 52 in relation to the basal ganglia. It lies in the anterior portion of the mesencephalon and contains neuronal cell bodies that secrete *dopamine* at their nerve endings. Additional neurons that secrete dopamine are also found in adjacent regions of the ventral tegmentum of the mesencephalon. In addition to the basal ganglia, fibers from these areas pass to the hypothalamus, the cerebral cortex, and the limbic system. The dopamine released at the nerve endings functions in the basal ganglia as an inhibitory transmitter, which is perhaps also true in other areas of the brain, but not necessarily so because some dopamine receptors are excitatory.

It will be recalled from Chapter 52 that destruction of the dopaminergic neurons in the substantia nigra eliminates their normal inhibitory effect on the basal ganglia, and that is the basic cause of *Parkinson's disease.*

(3) *The locus ceruleus.* This is a small area located bilaterally and posteriorly at the junction

point between the mesencephalon and the pons. Nerve fibers from this area secrete norepinephrine. They spread very widely through the diencephalon and the cerebrum, as well as into the cerebellum and lower regions of the brain stem. The norepinephrine released by the endings of these neurons may have either excitatory or inhibitory effects on different structures, depending on the character of the receptors in the postsynaptic neurons. Later in the chapter we shall see that the locus ceruleus probably plays a role in one type of sleep called REM sleep.

(4) *The raphe nuclei.* In the midline of the lower pons and medulla are several very thin nuclei called the raphe nuclei. Many of the neurons in these nuclei secrete serotonin, and they send fibers to widespread areas in the diencephalon and also to the cord. The cord fibers have the capability of suppressing pain, which was discussed in Chapter 50. The serotonin released in the diencephalon and cerebrum almost certainly plays an essential role in causing normal sleep, as we shall discuss later in the chapter.

THE GENERALIZED THALAMOCORTICAL SYSTEM—ITS POSSIBLE FUNCTION IN ATTENTION AND SEARCHING THE MEMORY STORE

The thalamus is the entryway for all sensory nervous signals to the cerebral cortex, with the single exception of signals from the olfactory system. Those thalamic nuclei that relay these sensory signals to the cerebral cortex are called the *specific* thalamic nuclei, and this sensory input system to the cerebral cortex is known as the *specific thalamocortical system*.

In addition to the *specific* thalamocortical system, another thalamocortical system separate from the specific system is called the *generalized thalamocortical system*. This is composed principally of the nerve pathway discussed earlier feeding upward from the mesencephalic portion of the reticular formation by way of the thalamus to the cortex. The signals of this pathway are relayed in the thalamus by multiple small diffuse neurons that lie mainly between the specific thalamic nuclei or on the outer surface of the thalamus. Many of these neurons make multiple connections with the specific thalamic nuclei, but especially they project very small fibers to all parts of the cerebral cortex.

Mechanism of Cortical Activation by the Generalized Thalamocortical System. The way in which the generalized thalamocortical system activates the cerebral cortex is entirely different from activation by the specific sensory systems. Some of the differences are the following.

1. Stimulation of a specific thalamic nucleus—such as the ventrobasal complex that transmits somatic signals to the somatic sensory cortex—activates the cortex within 1 to 2 milliseconds, whereas stimulation of the generalized system causes no activation for approximately the first 25 milliseconds. The activation level builds up over a period of many milliseconds. These differences are illustrated in Figure 55–3.

2. At the end of stimulation, the activation of the cortex by the specific nuclei dies away within another few milliseconds, whereas the activation by the generalized system sometimes continues as an "after-discharge" for as long as 30 seconds.

3. Signals from the specific nuclei to the cortex activate mainly layer IV of the cortex, as was explained in Chapter 49, whereas activation of the generalized thalamic system activates mainly layers I and II of the cortex. Since this latter activation is prolonged and because layers I and II are the loci of many of the dendrites of the deeper cortical neurons, it is supposed that the stimulation by the generalized thalamic system mainly causes partial depolarization of large numbers of dendrites near the surface of the cortex; this in turn causes a generalized increase in the degree of facilitation of the cortex. When the cortex is thus facilitated, specific signals that enter the cortex from other sources are exuberantly received.

4. Stimulation in the generalized thalamic system *facilitates* an *area* of several square centimeters in the cortex, while stimulation at a point in a specific thalamic nucleus *excites* a specific point in the cortex.

In summary, the generalized thalamocortical system controls the overall degree of activity of the cortex. It can at times facilitate activity in regional areas of the cortex distinct from the remainder of the cortex. Collateral signals from this system also control the level of activity in the specific nuclei of the thalamus, the basal ganglia, the hypothalamus, and other structures of the cerebrum and diencephalon as well.

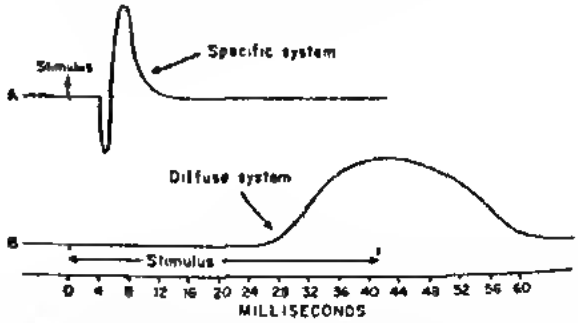


Figure 55–3. Comparison of the response of the visual cortex following stimulation of the lateral geniculate body (record A), which is part of the specific thalamocortical system, with the response of the visual cortex following stimulation of the visual portion of the generalized thalamocortical system (record B). Note in record B the long latent period before appearance of the diffuse response and also the prolonged duration of the response.

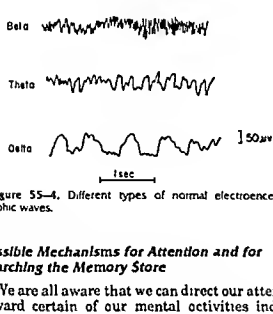


Figure 55–4. Different types of normal electroencephalographic waves.

Possible Mechanisms for Attention and for Searching the Memory Store

We are all aware that we can direct our attention toward certain of our mental activities individually and can also search through our memory store for specific memories. Because of the capability of the generalized thalamocortical system to activate small areas of the cerebral cortex at a time, it is tempting to believe that specific activation of regional portions of the cortex might be the way in which we do indeed direct our attention, and might also be the basis for searching through memory stores.

One other bit of information also suggests that the generalized thalamocortical system might be important in searching for memories: It has been reported that specific lesions in the thalamus are sometimes associated with retrograde amnesia—that is, inability to recall memories that are known to be stored within the brain.

BRAIN WAVES

Electrical recordings from the surface of the brain or from the outer surface of the head demonstrate continuous electrical activity in the brain. Both the intensity and patterns of this electrical activity are determined to a great extent by the overall level of excitation of the brain resulting from functions in the reticular activating system. The undulations in the recorded electrical potentials, shown in Figure 55–4, are called *brain waves*, and the entire record is called an *electroencephalogram* (EEG).

The intensities of the brain waves on the surface of the scalp range from 0 to 300 microvolts, and their frequencies range from once every few seconds to 50 or more per second. The character of the waves is highly dependent on the degree of activity of the cerebral cortex, and the waves change markedly between the states of wakefulness and sleep and coma.

Much of the time, the brain waves are irregular, and no general pattern can be discerned in the EEG. However, at other times, distinct patterns do appear. Some of these are characteristic of specific abnormalities of the brain, such as epilepsy, which is discussed later. Others occur even in normal persons and can be classified as *alpha, beta, theta,* and *delta waves*, which are all illustrated in Figure 55–4.

Alpha waves are rhythmic waves occurring at a frequency of between 8 and 13 per second and are found in the EEGs of almost all normal adult persons when they are awake in a quiet, resting state of cerebration. These waves occur most intensely in the occipital region but can also be recorded at times from the parietal and frontal regions of the scalp. Their voltage usually is about 50 microvolts. During sleep the alpha waves disappear entirely, and when the awake person's attention is directed to some specific type of mental activity, the alpha waves are replaced by asynchronous, higher frequency but lower voltage beta waves. Figure 55–5 illustrates the effect on the alpha waves of simply opening the eyes in bright light and then closing the eyes again. Note that the visual sensations cause immediate cessation of the alpha waves and that these are replaced by low voltage, asynchronous beta waves.

Beta waves occur at frequencies of more than 14 cycles per second and as high as 25 and rarely 50 cycles per second. These are most frequently recorded from the parietal and frontal regions of the scalp. Most beta waves appear during activation of the central nervous system or during tension.

Theta waves have frequencies of between 4 and 7 cycles per second. These occur mainly in the parietal and temporal regions in children, but they also occur during emotional stress in some adults, particularly during disappointment and frustration. They can often be brought out in the EEG of a frustrated person by allowing enjoyment of some pleasant experience and then suddenly removing this element of pleasure; this causes approximately 20 seconds of theta waves. These same waves also occur in many brain disorders.

Delta waves include all the waves of the EEG below 3.5 cycles per second and sometimes as low as 1 cycle every 2 to 3 seconds. These occur in deep sleep, in infancy, and in serious organic brain disease. And they occur in the cortex of animals that have had subcortical transections separating the cerebral cortex from the thalamus. Therefore, delta waves can occur strictly in the cortex independently of activities in lower regions of the brain.

ORIGIN OF THE DIFFERENT TYPES OF BRAIN WAVES

The discharge of a single neuron or single nerve fiber in the brain cannot be recorded from the surface of the head. Instead, for an electrical potential to be recorded all the way through the skull, large portions of nervous tissue must emit electrical current simultaneously. There are two ways in which this can occur. First, tremendous numbers of nerve fibers can discharge in synchrony with each other, thereby generating very strong electrical currents. Second, large numbers of neurons can partially discharge, though not emit action

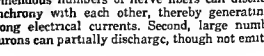


Figure 55–5. Replacement of the alpha rhythm by an asynchronous discharge on opening the eyes

potentials; furthermore, these partially discharged neurons can cause electrical potentials that undulate with changing degrees of excitability of the neurons. Simultaneous electrical measurements within the brain while recording brain waves from the scalp indicate that it is the second of these that causes the usual brain waves.

To be more specific, the surface of the cerebral cortex is composed almost entirely of a mat of dendrites extending to the surface from neuronal cells in the lower layers of the cortex. When signals impinge on these dendrites, the dendrites become partially discharged. This makes the neurons of the cortex highly excitable—that is, facilitates them, and the negative potential is simultaneously recorded from the surface of the scalp, indicating this high degree of excitability

One of the principal sources of signals to excite the outer dendritic layer of the cerebral cortex is the generalized thalamocortical fibers of the reticular activating system. Therefore, brain wave intensity is closely related to the degree of activity in either the brain stem or the thalamic portions of the reticular activating system.

Origin of Delta Waves. Transection of the fiber tracts from the thalamus to the cortex, which blocks the reticular activating system fibers, causes delta waves in the cortex This indicates that some synchronizing mechanism can occur in the cortical neurons themselves—entirely independently of lower structures in the brain—to cause the delta waves

Delta waves also occur in very deep "slow wave" sleep, and this suggests that the cortex then might be released from the activating influences of the reticular activating system.

Origin of Alpha Waves. Alpha waves will *not* occur in the cortex without connections with the thalamus. Also, stimulation in the generalized thalamic nuclei often sets up waves in the generalized thalamocortical system at a frequency of between 8 and 13 per second, the natural frequency of the alpha waves Therefore, it is assumed that the alpha waves result from spontaneous activity in the generalized thalamocortical system, which causes both the periodicity of the alpha waves and the synchronous activation of literally millions of cortical neurons during each wave.

EFFECT OF VARYING DEGREES OF CEREBRAL ACTIVITY ON THE BASIC RHYTHM OF THE ELECTROENCEPHALOGRAM

There is a general relationship between the degree of cerebral activity and the average frequency of the electroencephalographic rhythm, the average frequency increasing progressively with higher and higher degrees of activity This is illustrated in Figure 55–6, which shows the existence of delta waves in stupor, surgical anesthesia, and sleep; theta waves in psychomotor states and in infants; alpha waves during relaxed states; and beta waves during periods of intense mental activity. However, during periods of mental activity the waves usually become asynchronous rather than synchronous so that the voltage falls considerably, despite increased cortical activity, as illustrated in Figure 55–5

CLINICAL USE OF THE ELECTROENCEPHALOGRAM

One of the most important uses of the EEG is to diagnose different types of epilepsy and to find the focus in the brain causing the epilepsy. This is discussed further later. But, in addition, the EEG can be used to localize brain tumors or other space-occupying lesions of the brain and to diagnose certain types of psychopathic disturbances

Localization of Brain Tumors. There are two means by which brain tumors can be localized. Some brain tumors are so large that they block electrical activity from a given portion of the cerebral cortex, and when this occurs the voltage of the brain waves is considerably reduced in the region of the tumor. However, more frequently a brain tumor compresses the surrounding neuronal tissue and causes abnormal electrical excitation of these surrounding areas; this in turn leads to synchronous discharges of very high voltage waves in the EEG, as shown in the middle two records of Figure 55–7. Localization of the origin of these spikes on the surface of the scalp is a valuable means for locating the brain tumor; this is achieved in the following way.

The upper part of Figure 55–7 shows the placement of 16 different electrodes on the scalp, and the lower part of the figure shows the brain waves from 4 of these electrodes marked in the figure by Xs. Note that in two of these, intense brain waves are recorded and, furthermore, that the two waves are essentially of reverse polarity to each other This reverse polarity means that the origin of the spikes is somewhere in the area *between* the two respective electrodes. Thus, the excessively excitable area of the brain has been located, and this is a lead to the location of the brain tumor.

Diagnosing Psychopathic Disturbances. Use of brain waves in diagnosing psychopathic abnormalities is generally not very satisfactory because only a few of these cause distinct brain wave patterns. Yet by observing combinations of different types of basic rhythms, *reactions of the rhythms to attention, changes of the* rhythms during alkalosis caused by forced breathing, the appearance of particular characteristics in the brain waves (such as "spindles" of alpha waves), and so forth, an experienced electroencephalographer can detect at least certain types of psychopathic disturbances. Also, theta waves are frequently found in persons with brain abnormalities.

SLEEP

Sleep is defined as a state of unconsciousness from which a person can be aroused by appropriate sensory or other stimuli.

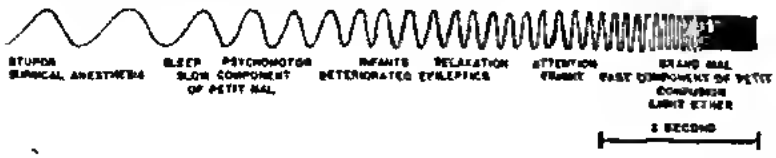


Figure 55–6. Effect of varying degrees of cerebral activity on the basic rhythm of the EEG. (From Gibbs and Gibbs. Atlas of Electroencephalography. 2nd Ed. Vol. I. Addison-Wesley, 1974 Reprinted by permission.)

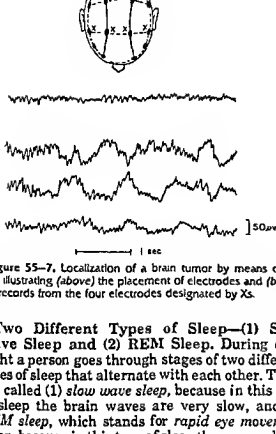


Figure 55–7. Localization of a brain tumor by means of the EEG, illustrating *(above)* the placement of electrodes and *(below)* the records from the four electrodes designated by Xs.

Two Different Types of Sleep—(1) Slow Wave Sleep and (2) REM Sleep. During each night a person goes through stages of two different types of sleep that alternate with each other. These are called (1) *slow wave sleep,* because in this type of sleep the brain waves are very slow, and (2) *REM sleep,* which stands for *rapid eye movement* sleep, because in this type of sleep the eyes undergo rapid movements despite the fact that the person is still asleep.

Most of the sleep during each night is of the slow wave variety; this is the deep, restful type of sleep that the person experiences during the first hour of sleep after having been kept awake for many hours. Episodes of REM sleep occur periodically during sleep and occupy about 25 per cent of the sleep time of the young adult; they normally recur about every 90 minutes. This type of sleep is not so restful, and it is usually associated with dreaming, as we shall discuss later.

Electroencephalographic Changes in the Different Stages of Wakefulness and Sleep

Figure 55–8 illustrates the electroencephalogram from a typical person in different stages of wakefulness and sleep. Alert wakefulness is characterized by high frequency *beta waves,* whereas quiet wakefulness is usually associated with *alpha waves,* as illustrated by the first two electroencephalograms of the figure.

Slow wave sleep is generally divided into four stages. In the first stage, a stage of very light sleep, the voltage of the electroencephalographic waves becomes very low but this is broken by "*sleep spindles,*" that is, short spindle-shaped bursts of alpha waves that occur periodically. In stages 2, 3, and 4 of slow wave sleep the frequency of the electroencephalogram becomes progressively slower until it reaches a frequency of only two to three waves per second; these are typical *delta waves.*

In REM sleep, the electroencephalogram suddenly changes back to the characteristics of the early stages of wakefulness, indicating a high level of activity in the brain during this period of sleep.

SLOW WAVE SLEEP

Most of us can understand the characteristics of deep slow wave sleep by remembering the last time that we were kept awake for more than 24 hours and the deep sleep that occurred within 30 minutes to an hour after going to sleep. This sleep is exceedingly restful and is associated with a decrease in both peripheral vascular tone and many other vegetative functions of the body as well. In addition, there is a 10 to 30 per cent decrease in blood pressure, respiratory rate, and basal metabolic rate.

Though slow wave sleep is frequently called "dreamless sleep," dreams actually occur very often during slow wave sleep, and nightmares even occur during this type of sleep. However, the difference between the dreams occurring in slow wave sleep and those in REM sleep is that those of REM sleep are remembered whereas those of slow wave sleep usually are not. That is, during slow wave sleep the process of consolidation of the dreams in memory does not occur.

Origin of the Delta Waves in Sleep. In experimental animals in which the fiber tracts between the thalamus and the cortex have been transected, delta waves are generated in the isolated cortex, indicating that this type of wave probably occurs intrinsically in the cortex when the cortex is not driven from below by the reticular activating system. Therefore, it is assumed that the degree of

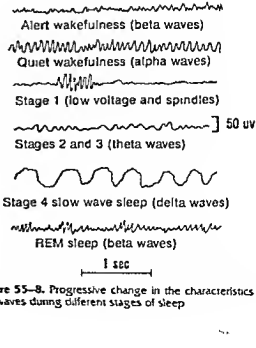


Figure 55–8. Progressive change in the characteristics of the brain waves during different stages of sleep

activity in the activating system has fallen to a level during deep slow wave sleep that is too low to maintain normal excitability of the cortex.

REM SLEEP (PARADOXICAL SLEEP, DESYNCHRONIZED SLEEP)

In a normal night of sleep, bouts of REM sleep lasting 5 to 30 minutes usually appear on the average every 90 minutes, the first such period occurring 80 to 100 minutes after the person falls asleep. When the person is extremely tired, the duration of each bout of REM sleep is very short, and it may even be absent. On the other hand, as the person becomes more rested through the night, the duration of the REM bouts greatly increases.

There are several very important characteristics of REM sleep:

1. It is usually associated with active dreaming
2. The person is even more difficult to arouse by sensory stimuli than during deep slow wave sleep, and yet persons usually awaken in the morning during an episode of REM sleep, not from slow wave sleep.
3. The muscle tone throughout the body is exceedingly depressed, indicating strong inhibition of the spinal projections from the reticular formation of the brain stem
4. The heart rate and respiration usually become irregular, which is characteristic of the dream state.
5. Despite the extreme inhibition of the peripheral muscles, a few irregular muscle movements occur These include, in particular, rapid movements of the eyes; this is the origin of the name acronym REM, for "rapid eye movements"
6. The electroencephalogram shows a desynchronized pattern of beta waves similar to those that occur during wakefulness Therefore, this type of sleep is also frequently called *desynchronized sleep*, meaning desynchronized brain waves, or, it is also called *paradoxical sleep* because it is a paradox that a person can still be asleep despite marked activity in the brain

In summary, REM sleep is a type of sleep in which the brain is quite active However, the brain activity is not channeled in the proper direction for persons to be aware of their surroundings and therefore to be awake.

BASIC THEORIES OF SLEEP

The Active Theory of Sleep. An earlier theory of sleep was that the reticular activating system and other parts of the brain simply fatigued over the period of a waking day and therefore became inactive as a result. This was called the *passive theory of sleep*. However, an important experiment quoted earlier in this chapter changed this view to the current belief that sleep most likely is caused by an active inhibitory process. This was the experiment in which it was discovered that transecting the brain stem in the midpontile region leads to a brain that never goes to sleep. In other words, there seems to be some center or centers located below the midpontile level of the brain stem that actively cause sleep by inhibiting other parts of the brain. This is called the *active theory* of sleep.

Neuronal Centers, Transmitters, and Mechanisms That Can Cause Sleep

Stimulation of several specific areas of the brain can produce sleep with characteristics very near those of natural sleep. Some of these are the following.

1 The most conspicuous stimulation area for causing almost natural sleep is the raphe nuclei in the lower half of the pons and in the medulla. These are a thin sheet of nuclei located in the midline. Nerve fibers from these nuclei spread widely in the reticular formation and also upward into the thalamus and neocortex, hypothalamus, and most areas of the limbic cortex In addition, they extend downward into the spinal cord, terminating in the posterior horns where they can inhibit incoming pain signals, as was discussed in Chapter 50 It is also known that the endings of fibers from these raphe neurons secrete *serotonin* Therefore, it is assumed that serotonin is the major transmitter substance associated with production of sleep

2 Stimulation of some areas in the *nucleus of the tractus solitarius*, which is the sensory region of the medulla and pons for the visceral sensory signals entering the brain via the vagi and glossopharyngeal nerves, will also promote sleep However, this will not occur if the raphe nuclei have been destroyed. Therefore, these regions probably act by exciting the raphe nuclei.

3. Stimulation of several other regions in the lower brain stem and diencephalon can also help promote sleep, including (a) the rostral part of the hypothalamus, mainly in the suprachiasmal area and (b) an occasional area in the diffuse nuclei of the thalamus.

Effect of Lesions in the Sleep-Promoting Centers. Discrete lesions in the raphe nuclei lead to a high state of wakefulness. This is also true of bilateral lesions in the mediorostral suprachiasmal portion of the anterior hypothalamus. In both instances, the reticular activating system seems to become released from inhibition. Indeed, the lesions of the anterior hypothalamus can sometimes cause such intense wakefulness that the animal actually dies of exhaustion

Other Possible Transmitter Substances Related to Sleep. Experiments have shown that the cerebrospinal fluid and also the blood of animals that have been kept awake for several days contain a substance or substances that cause sleep when

injected into the ventricular system of an animal. One of these substances is a small polypeptide with a molecular weight of less than 500. When cerebrospinal fluid containing this sleep-producing substance or substances is injected into the third ventricle, almost natural sleep occurs within a few minutes, and the animal may then stay asleep for several hours. Another substance that has similar effects in causing sleep is a nonapeptide isolated from the blood of sleeping animals And still a third and different sleep factor has been isolated from the neuronal tissues of the brain stem of animals kept awake for days. Therefore, it is possible that prolonged wakefulness causes progressive accumulation of a sleep factor in the brain stem or in the cerebrospinal fluid that leads to sleep.

Possible Causes of REM Sleep

Why slow wave sleep is broken periodically by REM sleep is not understood However, a lesion in the *locus ceruleus* on each side of the brain stem can reduce REM sleep, and if the lesion includes other contiguous areas of the brain stem, REM sleep can be prevented altogether Therefore, it has been postulated that when stimulated, the norepinephrine-secreting nerve fibers that originate in the locus ceruleus can activate many portions of the brain This theoretically causes the excess activity that occurs in certain regions of the brain in REM sleep, but the signals are not channeled appropriately in the brain to cause normal conscious awareness that is characteristic of the wakefulness state.

The Cycle Between Sleep and Wakefulness

The preceding discussions have merely identified neuronal areas, transmitters, and mechanisms that are related to sleep. However, they have not explained the cyclic, reciprocal operation of the sleep-wakefulness cycle. It is quite possible that this is caused by a free-running intrinsic oscillator within the brain stem that cycles back and forth between the sleep centers and areas in the reticular activating system that cause wakefulness.

However, there is much reason to believe that feedback signals from the cerebral cortex and also from the peripheral nerve receptors might also play an important role in helping control the sleep-wakefulness rhythm. One reason for believing this is that sensory signals feeding into the reticular activating system will often arouse a person from deep sleep. Also, as explained earlier, stimulation of the cerebral cortex will cause powerful activation of the reticular activating system.

Therefore, a very likely mechanism for causing the rhythmicity of the sleep-wakefulness cycle is the following:

When the sleep centers are not activated, the reticular activating system then presumably begins spontaneous activity. This in turn excites both the cerebral cortex and the peripheral nervous system. Next, positive feedback signals come from both these areas back to the reticular activating system to activate it still further. Thus, once the wakefulness state begins, it has a natural tendency to sustain itself.

However, after the brain remains activated for many hours, even the neurons within the activating system presumably will fatigue to some extent, and other factors presumably activate the sleep centers. Consequently, the positive feedback cycle between the reticular activating system and the cortex, and also that between the reticular activating system and the periphery, will fade, and the inhibitory effects of the sleep centers as well as inhibition by possible sleep-producing chemical transmitter substances will take over, leading to rapid transition from the wakefulness state to the sleep state

Then, one could postulate that during sleep the excitatory neurons of the reticular activating system gradually become more and more excitable because of the prolonged rest, while the inhibitory neurons of the sleep centers become less excitable, thus leading to a new cycle of wakefulness.

This theory obviously can explain the rapid transitions from sleep to wakefulness and from wakefulness to sleep It can also explain arousal, the insomnia that occurs when a person's mind becomes preoccupied with a thought, the wakefulness that is produced by bodily activity, and many other conditions that affect the person's state of sleep or wakefulness.

PHYSIOLOGICAL EFFECTS OF SLEEP

Sleep causes two major types of physiological effects: first, effects on the nervous system itself and, second, effects on other structures of the body. The first of these seems to be by far the more important, for any person who has a transected spinal cord in the neck shows no physiological effects in the body beneath the level of transection that can be attributed to a sleep and wakefulness cycle; that is, lack of sleep and wakefulness causes neither significant harm to the bodily organs nor even any deranged function. On the other hand, lack of sleep certainly does affect the functions of the central nervous system.

Prolonged wakefulness is often associated with progressive malfunction of the mind and also causes abnormal behavioral activities of the nervous system. We are all familiar with the increased sluggishness of thought that occurs toward the end of a prolonged wakeful period, but in addition, a person can become irritable or even psychotic following forced wakefulness for prolonged periods of time. Therefore, one can assume that sleep in some way not presently understood restores both normal

sensitivities of and normal "balance" among the different parts of the central nervous system. This might be likened to the "rezeroing" of electronic analog romputers after prolonged use, for all computers of this type gradually lose their "base line" of operation; it is reasonable to assume that the same effect occurs in the central nervous system, because overuse of some neurons during wakefulness could easily throw all these out of balance with the remainder of the nervous system. Therefore, in the absence of any definitely demonstrated functional value of sleep, we might postulate that the principal value of sleep is to restore the natural balance among the neuronal centers

Even though, as pointed out earlier, wakefulness and sleep have not been shown to be necessary for somatic functions of the body, the cycle of enhanced and depressed nervous excitability that follows the cycle of wakefulness and sleep does have moderate effects on the peripheral body. For instance, there is enhanced sympathetic activity during wakefulness and also enhanced numbers of impulses to the skeletal musculature to increase muscle tone Conversely, during sleep, sympathetic activity decreases while parasympathetic activity occasionally increases Therefore, arterial blood pressure falls, pulse rate decreases, skin vessels dilate, activity of the gastrointestinal tract sometimes increases, muscles fall into a mainly relaxed state, and the overall basal metabolic rate of the body falls by 10 to 30 per cent

EPILEPSY

Epilepsy is characterized by uncontrolled excessive activity of either a part or all of the central nervous system A person who is predisposed to epilepsy has attacks when the basal level of excitability of the nervous system (or of the part that is susceptible to the epileptic state) rises above a certain critical threshold But, as long as the degree of excitability is held below this threshold, no attack occurs

Basically, epilepsy can be classified into three major types: *grand mal epilepsy*, *petit mal epilepsy*, and *focal epilepsy*

GRAND MAL EPILEPSY

Grand mal epilepsy is characterized by extreme neuronal discharges in all areas of the brain—in the cortex, in the deeper parts of the cerebrum, and even in all areas of the reticular activating system. Also, discharges are transmitted from the reticular formation into the spinal cord to cause generalized *tonic convulsions* of the entire body, followed toward the end of the attack by alternating muscular contractions called *tonic-clonic convulsions* Often the person bites or "swallows" the tongue and usually has difficulty in breathing, sometimes to the extent of developing cyanosis. Also, signals to the viscera frequently cause urination and defecation

The grand mal seizure lasts from a few seconds to as long as three to four minutes and is characterized by depression of the entire nervous system; the remains in stupor for one to many minutes after

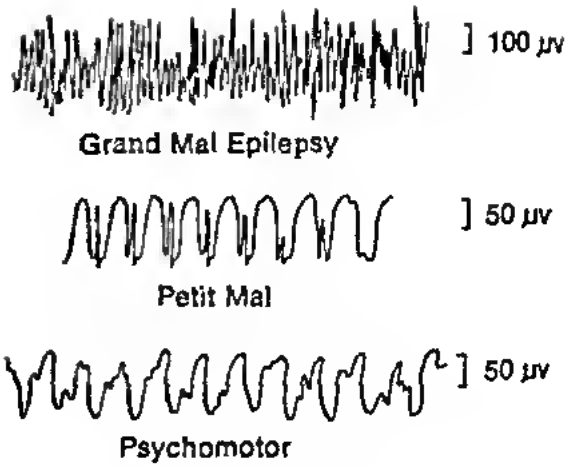


Figure 55–9. Electroencephalograms in different types of epilepsy

the attack is over and then often remains severely fatigued for many hours thereafter

The top recording of Figure 55–9 illustrates a typical electroencephalogram from almost any region of the cortex during the tonic phase of a grand mal attack. This illustrates that high voltage, synchronous discharges occur over the entire cortex. Furthermore, the same type of discharge occurs on both sides of the brain at the same time, illustrating that the abnormal neuronal circuitry responsible for the attack strongly involves the basal regions of the brain that drive the cortex

In experimental animals or even in human beings, grand mal attacks can be initiated by administering neuronal stimulants, such as the well-known drug Metrazol, or they can be caused by insulin hypoglycemia or by the passage of alternating electrical current directly through the brain. Electrical recordings from the thalamus and also from the reticular formation of the brain stem during the grand mal attack show typical high voltage activity in both of these areas similar to that recorded from the cerebral cortex.

Presumably, therefore, a grand mal attack is caused by abnormal activation in the lower portions of the brain activating system itself.

What Initiates a Grand Mal Attack? Most persons who have grand mal attacks have a hereditary predisposition to epilepsy, a predisposition that occurs in about 1 of every 50 to 100 people. In such persons, some of the factors that can increase the excitability of the abnormal "epileptogenic" circuitry enough to precipitate attacks are (1) strong emotional stimuli, (2) alkalosis caused by overbreathing, (3) drugs, (4) fever, or (5) loud noises or flashing lights Also, even in persons not genetically predisposed, traumatic lesions in almost any part of the brain can cause excess excitability of local brain areas as we shall discuss shortly, and these too can transmit signals into the reticular activating system to elicit grand mal seizures

What Stops the Grand Mal Attack? The cause of the extreme neuronal overactivity during a grand mal attack is presumed to be massive activation of many reverberating pathways throughout the brain. It is presumed that the major factor, or at least one of the major factors, that stops the attack after a few minutes is the phenomenon of neuronal *fatigue*. However, a second factor is probably *active inhibition* by certain structures of the brain. The stupor and total body fatigue that

occur after a grand mal seizure is over are believed to result from the intense fatigue of the neurons following their intensive activity during the grand mal attack.

PETIT MAL EPILEPSY

Petit mal epilepsy is closely allied to grand mal epilepsy in that it too almost certainly involves the reticular activating system. It is usually characterized by 3 to 30 seconds of unconsciousness during which the person has several twitchlike contractions of the muscles, usually in the head region—especially blinking of the eyes; this is followed by return of consciousness and resumption of previous activities. The patient may have one such attack in many months or in rare instances may have a rapid series of attacks, one following the other. However, the usual course is for the petit mal attacks to appear in late childhood and then to disappear entirely by the age of 30. On occasion, a petit mal epileptic attack will initiate a grand mal attack.

The brain wave pattern in petit mal epilepsy is illustrated by the middle record of Figure 55–9, which is typified by a *spike and dome pattern*. The spike portion of this recording is almost identical to the spikes that occur in grand mal epilepsy, but the dome portion is distinctly different. The spike and dome can be recorded over most or all of the cerebral cortex, illustrating that the seizure involves the activating system of the brain

FOCAL EPILEPSY

Focal epilepsy can involve almost any part of the brain, either localized regions of the cerebral cortex or deeper structures of both the cerebrum and brain stem. And almost always, focal epilepsy results from some localized organic lesion or functional abnormality, such as a scar that pulls on the neuronal tissue, a tumor that compresses an area of the brain, a destroyed area of brain tissue, or congenitally deranged local circuitry. Lesions such as these can promote extremely rapid discharges in the local neurons, and when the discharge rate rises above approximately 1000 per second, synchronous waves begin to spread over the adjacent cortical regions. These waves presumably result from *localized reverberating circuits* that gradually recruit adjacent areas of the cortex into the discharge zone. The process spreads to adjacent areas at a rate as slow as a few millimeters a minute to as fast as several centimeters per second. When such a wave of excitation spreads over the motor cortex, it causes a progressive "march" of muscular contractions throughout the opposite side of the body, beginning most characteristically in the mouth region and marching progressively downward to the legs, but at other times marching in the opposite direction. This is called *jacksonian epilepsy*.

A focal epileptic attack may remain confined to a single area of the brain, but in many instances the strong signals from the convulsing cortex or other part of the brain excite the mesenrephalic portion of the brain activating system so greatly that a grand mal epileptic attack ensues as well.

Another type of focal epilepsy is the so-called *psychomotor seizure*, which may cause (1) a short period of amnesia, (2) an attack of abnormal rage, (3) sudden anxiety, discomfort, or fear, (4) a moment of incoherent speech or mumbling of some trite phrase, or (5) a motor act to attack someone, to rub the face with the hand, or so forth. Sometimes the person cannot remember his activities during the attack, but at other times he will have been conscious of everything that he had been doing but unable to control it. Attacks of this type characteristically involve part of the limbic portion of the brain, such as the hippocampus, the amygdala, the septum, or the temporal cortex.

The lower tracing of Figure 55–9 illustrates a typical electroencephalogram during a psychomotor attack, showing a low frequency rectangular wave with a frequency of between 2 and 4 per second and with superimposed 14 per second waves.

The electroencephalogram can often be used to localize abnormal spiking waves originating in areas of organic brain disease that might predispose to focal epileptic attacks Once such a focal point is found, surgical excision of the focus frequently prevents future attacks.

REFERENCES

Arkin, J S, *et al* (eds) The Mind in Sleep New York, Halsted Press, 1978

Bindman, L The Neurophysiology of the Cerebral Cortex Austin, University of Texas Press, 1981

Block, G D, and Fagg, T L Circadian pacemakers in the nervous system *Annu Rev Neurosci*, *1* 19, 1978

Buser, P Higher functions of the nervous system *Annu Rev Physiol*, 38 217, 1978

Enright J T The Timing of Sleep and Wakefulness On the Substructure and Dynamics of the Circadian Pacemakers Underlying the Wake-Sleep Cycle New York, Springer-Verlag, 1979

Fuller, R W Pharmacology of central serotonin neurons *Annu Rev Pharmacol Toxicol*, *20* 111, 1980

Glaser, G H, *et al* (eds) Antiepileptic Drugs New York, Raven Press, 1980

Hobson, J A, and Brazier, M A B (eds) The Reticular Formation Revisited Specifying Function for a Nonspecific System. New York, Raven Press, 1980

Hyvarinen, J Posterior parietal lobe of the primate brain *Physiol Rev*, 62 1060, 1982

Ito, M, *et al* (eds) Integrative Control Functions of the Brain New York, Elsevier/North-Holland, 1978

Jones, E G Organization of the thalamocortical complex and its relation to sensory processes *In* Darian-Smith, I. (ed) Handbook of Physiology Sec 1, Vol III. Bethesda, American Physiological Society, 1984, p 149

Klee, M (ed) Physiology and Pharmacology of Epileptogenic Phenomena New York, Raven Press, 1982

Livingston, R B, Sensory Processing, Perception, and Behavior New York, Raven Press, 1978

Mitzdorf, U: Current source-density method and application in cat cerebral cortex: Investigation of evoked potentials and EEG phenomena *Physiol Rev*, 65 37, 1985

Newmark, M E, and Penry, J K Genetics of Epilepsy A Review New York, Churchill Livingstone, 1980

O Keefe, J, and Nadel, L. The Hippocampus as a Cognitive Map New York, Oxford University Press, 1978

Pappenheimer, J R. The sleep factor. *Sci Am*, *235*(2) 24, 1976

Passouant, P., and Oswald, I. (eds) Pharmacology of the States of Alertness. New York, Pergamon Press, 1979

Plum, F, and Posner, J B The Diagnosis of Stupor and Coma, 3rd Ed. Philadelphia, F. A. Davis, 1980

Pollack, M H, *et al*: Propranolol and depression revisited. Three cases and a review *J. Nerv. Ment Dis*, *173*.118, 1985

Regan, D: Electrical responses evoked from the human brain *Sci Am*, *241*(6)-134, 1979.

Schneider, A. M., and Tarshis, B.: An Introduction to Physiological Psychology. New York, Random House, 1979

Stern, R. M., *et al*: Psychophysiological Recording New York, Oxford University Press, 1980

Tucek, S.. Regulation of acetylcholine synthesis in the brain *J. Neurochem*, *44* 11, 1985.

56

Behavioral Functions of the Brain: The Limbic System, Role of the Hypothalamus, and Control of Vegetative Functions of the Body

Behavior is a function of the entire nervous system, not of any particular portion Even the discrete cord reflexes are an element of behavior, and the wakefulness and sleep cycle discussed in Chapter 55 is certainly one of the most important of our behavioral patterns However, in this chapter we will deal with those special types of behavior associated with emotions, subconscious motor and sensory drives, and the intrinsic feelings of punishment and pleasure These functions of the nervous system are performed mainly by subcortical structures located in the basal regions of the brain.

The word "limbic" means "border," and the original usage of the term *limbic system* was to describe the brain structures that lie in the border region between the hypothalamus and its related structures, on the one hand, and the cerebral cortex. However, as we have learned more about the functions of the hypothalamus and the limbic system, it has become clear that they mainly function together as a total system. Therefore, the term limbic system has now been expanded in common usage to mean this entire basal system of the brain that mainly controls the person's emotional behavior and drive.

The hypothalamus and its related structures also control many internal conditions of the body as well as aspects of behavior—such conditions as body temperature, osmolality of the body fluids, the drives to eat and drink, body weight, and so forth. These internal functions are collectively called *vegetative functions* of the body, and th control is obviously closely related to behavior.

FUNCTIONAL ANATOMY OF THE LIMBIC SYSTEM; ITS RELATION TO THE HYPOTHALAMUS

Figure 56–1 illustrates the anatomical structures the limbic system and its relationship to the hypoth amus, showing these to be an interconnected complex basal brain elements Located in the midst of all the is the hypothalamus, which is considered by ma anatomists to be a separate structure from the remai der of the limbic system but which, from a physiologic point of view, is one of the central elements of t system Therefore, Figure 56–2 illustrates schematical this key position of the hypothalamus in the limb system and shows that surrounding it are the oth subcortical structures of the limbic system, includir the *septum*, the *paraolfactory area*, the *epithalamus*, t *anterior nuclei of the thalamus*, *portions of the bas ganglia*, the *hippocampus*, and the *amygdala*

Surrounding the subcortical limbic areas is the *limb cortex* composed of a ring of cerebral cortex (a) beginnin in the *orbitofrontal area* on the ventral surface of th frontal lobes, (b) extending upward in the *subcallosa gyrus* beneath the anterior limb of the corpus callosum (c) over the top of the corpus callosum onto the media aspect of the cerebral hemisphere in the *cingulate gyrus* and finally (d) passing behind the corpus callosum an downward onto the ventromedial surface of the tempora lobe to the *parahippocampal gyrus* and *uncus*. Thus, o

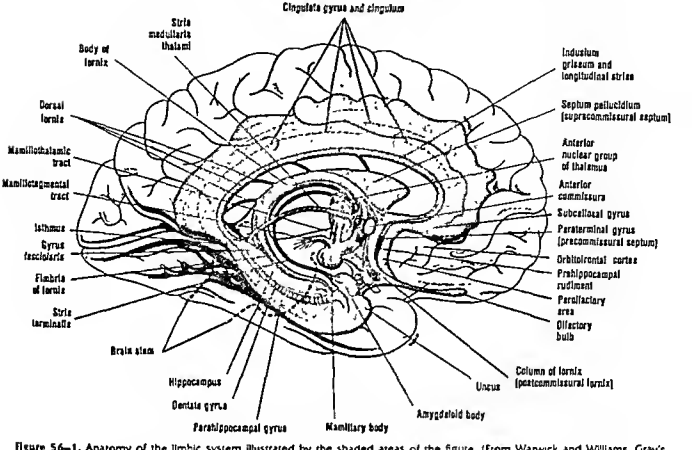


Figure 56–1. Anatomy of the limbic system illustrated by the shaded areas of the figure (From Warwick and Williams. Gray's Anatomy 35th Br ed London, Longman Group Ltd, 1973)

the medial and ventral surfaces of each cerebral hemisphere is a ring of mostly *paleocortex* that surrounds a group of deep structures intimately associated with overall behavior and with emotions In turn, this ring of limbic cortex functions as a two-way communication and association linkage between the *neocortex* and the lower limbic structures

It is also important to recognize that many of the behavioral functions elicited from the hypothalamus and from other limbic structures are mediated through the reticular formation of the brain stem. It was pointed out

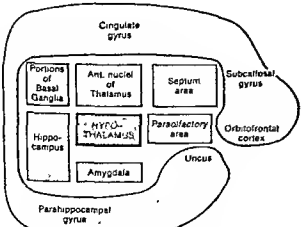


Figure 56–2. The limbic system.

in Chapters 52 and 55 that stimulation of the excitatory portion of the reticular formation can cause high degrees of somatic excitability, in Chapter 57 we will see that most of the signals for control of the autonomic nervous system also are transmitted through nuclei located in the brain stem

Therefore, from a physiological point of view the reticular formation and allied regions of the brain stem are a very important part of the limbic system even though anatomically they are considered to be a separate entity A very important route of communication between the limbic system and the brain stem is the *medial forebrain bundle* that extends from the septal *and orbitofrontal cortical regions* downward through the hypothalamus to the reticular formation This bundle carries fibers in both directions, forming a trunk-line communication system A second route of communication is through short pathways among the reticular formation, the thalamus, the hypothalamus, and most of the other contiguous areas of the basal brain.

THE HYPOTHALAMUS, A MAJOR OUTPUT PATHWAY OF THE LIMBIC SYSTEM

Note in Figure 56–2 that the hypothalamus lies in the very middle of the limbic system. It also has communicating pathways with all levels of this system. In turn, it and its closely allied structures send output signals in three directions. (1) down-

ward through the brain stem mainly into the reticular formation of the mesencephalon, pons, and medulla, (2) upward toward many higher areas of the diencephalon and cerebrum, especially the anterior thalamus and the limbic cortex, and (3) into the infundibulum to control most of the secretory functions of both the posterior and anterior pituitary glands.

In summary, the hypothalamus, which represents less than 1 per cent of the brain mass, nevertheless is the most important of all the motor output pathways of the limbic system. It controls most of the vegetative and endocrine functions of the body as well as many aspects of emotional behavior. Let us discuss first the vegetative and endocrine control functions of the hypothalamus and then return to its behavioral functions to see how all these operate together.

VEGETATIVE AND ENDOCRINE CONTROL FUNCTIONS OF THE HYPOTHALAMUS

The different hypothalamic mechanisms for controlling the vegetative and endocrine functions of the body are discussed along with the functions that they control in many different chapters throughout this text. For instance, the role of the hypothalamus in arterial pressure regulation is discussed in Chapter 21, thirst and water conservation in Chapter 36, temperature regulation in Chapter 72, and endocrine control in Chapter 74. However, to illustrate the organization of the hypothalamus as a functional unit, let us summarize the more important of its vegetative and endocrine functions here as well.

Figure 56–3 shows an enlargement of the hypothalamus, which represented only a small area in Figure 56–1. Please take a few minutes to study this diagram, especially to read the multiple activities that are excited or inhibited when respective hypothalamic nuclei are stimulated. In addition to those centers that are illustrated, a large *lateral hypothalamic* area overlies the illustrated area on each side of the hypothalamus. The lateral areas are especially important in controlling thirst, hunger, and many of the emotional drives.

A word of caution must be issued for studying this diagram, however, for the areas that cause specific activities are not nearly so discrete nor so accurately localized as suggested in the figure. Also, it is not known whether the effects noted in the figure result from stimulation of specific control nuclei or whether they result merely from activation of fiber tracts leading from control nuclei located elsewhere. For instance, stimulation of either the lateral hypothalamus or the posterior hypothalamus, through which many fiber pathways pass, can at times elicit many functions believed to be controlled primarily by other hypothalamic nuclei or even by nuclei outside the hypothalamus. With this caution in mind, we can give the following general description of the vegetative and control functions of the hypothalamus.

Cardiovascular Regulation. Stimulation of different areas throughout the hypothalamus can cause every known type of neurogenic effect on the cardiovascular system, including increased arterial pressure, decreased arterial pressure, increased heart rate, and decreased heart rate. In general, stimulation in the *posterior and lateral hypothalamus* increases the arterial pressure and

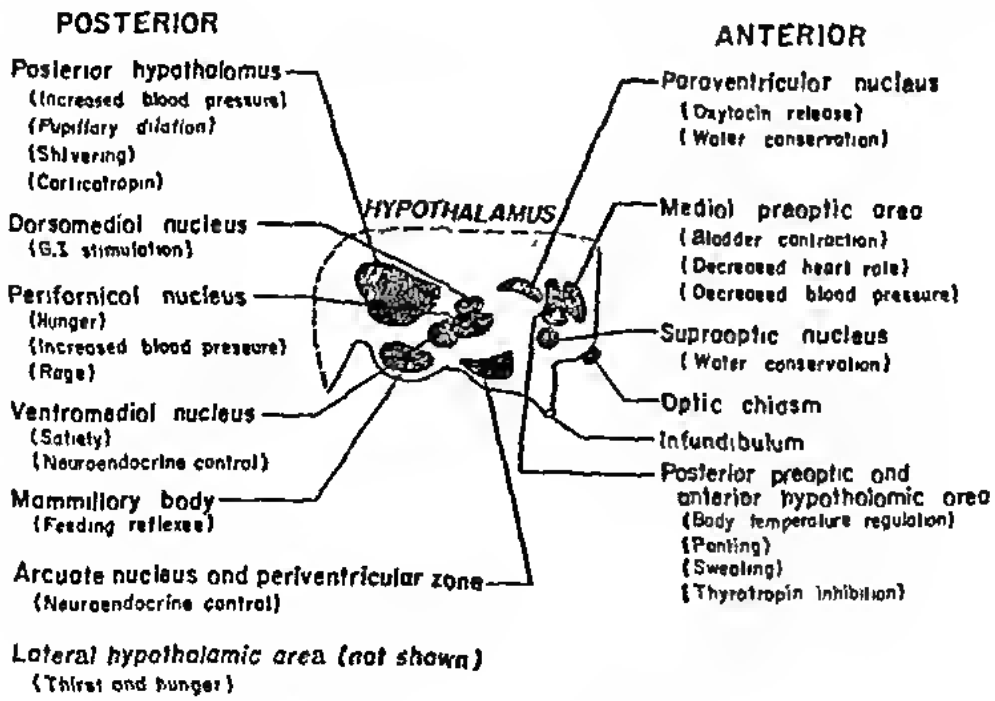


Figure 56–3. Control centers of the hypothalamus.

heart rate, while stimulation in the *preoptic area* often has opposite effects, causing a decrease in both heart rate and arterial pressure. These effects are transmitted mainly through the cardiovascular control centers in the reticular substance of the medulla and pons.

Regulation of Body Temperature. Large areas in the anterior portion of the hypothalamus, especially in the *preoptic area*, are concerned with regulation of body temperature. An increase in the temperature of the blood flowing through these areas increases the activity of temperature-sensitive neurons, whereas a decrease in temperature decreases their activity. In turn, these neurons control the mechanisms for increasing or decreasing body temperature, as discussed in Chapter 72.

Regulation of Body Water. The hypothalamus regulates body water in two separate ways. (1) by creating the sensation of thirst, which makes an animal drink water, and (2) by controlling the excretion of water into the urine. An area called the *thirst center* is located in the lateral hypothalamus When the electrolytes inside the neurons either of this center or allied areas of the hypothalamus become too concentrated, the animal develops an intense desire to drink water; it will search out the nearest source of water and drink enough to return the electrolyte concentration of the thirst center neurons to normal.

Control of renal excretion of water is vested mainly in the *supraoptic* nucleus. When the body fluids become too concentrated, the neurons of this area become stimulated The nerve fibers from these neurons project downward through the infundibulum into the posterior pituitary gland where they secrete a hormone called *antidiuretic hormone* (also called *vasopressin*) This hormone is then absorbed into the blood and acts on the collecting ducts of the kidneys to cause massive reabsorption of water, thereby decreasing the loss of water into the urine. These functions were presented in Chapter 36.

Regulation of Uterine Contractility and of Milk Ejection by the Breasts. Stimulation of the *paraventricular nuclei* causes its neuronal cells to secrete the hormone *oxytocin*. This in turn causes increased contractility of the uterus and also contraction of the myoepithelial cells that surround the alveoli of the breasts, which then causes the alveoli to empty the milk through the nipples. At the end of pregnancy, especially large quantities of oxytocin are secreted, and this secretion helps to promote labor contractions that expel the baby. Also, when a baby suckles the mother's breast, a reflex signal from the nipple to the hypothalamus causes oxytocin release, and the oxytocin then performs the necessary function of expelling the milk through the nipples so that the baby can nourish itself. These functions are discussed in Chapter 82.

Gastrointestinal and Feeding Regulation. Stimulation of several areas of the hypothalamus causes an animal to experience extreme hunger, a voracious appetite, and an intense desire to search for food. The area most associated with hunger is the *lateral hypothalamic area*. On the other hand, damage to this causes the animal to lose desire for food, sometimes causing lethal starvation.

A center that opposes the desire for food, called the *satiety center*, is located in the *ventromedial nucleus*. When this center is stimulated, an animal that is eating food suddenly stops eating and shows complete indifference to the food. On the other hand, if this area is destroyed, the animal cannot be satiated, but instead, its hypothalamic hunger centers become overactive so that it has a voracious appetite, resulting in tremendous obesity.

Another area of the hypothalamus that enters into the overall control of gastrointestinal activity is the *mammillary bodies* that activate many feeding reflexes, such as licking the lips and swallowing.

Hypothalamic Control of the Anterior Pituitary Gland

Stimulation of certain areas of the hypothalamus also causes the anterior pituitary gland to secrete its hormones. This subject will be discussed in detail in Chapter 74 in relation to the neural control of the endocrine glands, but, briefly, the basic mechanism of the control of the anterior pituitary is the following:

The anterior pituitary gland receives its blood supply mainly from veins that flow into the anterior pituitary sinuses from the lower part of the hypothalamus. As the blood courses through the hypothalamus before reaching the anterior pituitary, *releasing hormones* and *inhibitory hormones* are secreted into the blood by various hypothalamic nuclei They are then transported in the blood to the anterior pituitary where they act on the glandular cells to control the release of the anterior pituitary hormones.

The cell bodies of the neurons that secrete these releasing hormones and inhibitory hormones are located mainly in the medial basal nuclei of the hypothalamus, especially in the *paroventricular zone*, the *arcuate nucleus*, and part of the *ventromedial nucleus*. However, the axons from these nuclei then project to the *median eminence*, which is an enlarged area of the infundibulum where it arises from the inferior border of the hypothalamus It is here that the nerve terminals actually secrete their releasing and inhibitory hormones. These hormones are then absorbed into the blood capillaries in the median eminence and carried in the venous blood down along the infundibulum to the anterior pituitary gland

Summary. A number of discrete areas of the hypothalamus have now been found that control specific vegetative functions. However, these areas are still poorly delimited, so much so that the above separation of different areas for different hypothalamic functions is partially artificial

BEHAVIORAL FUNCTIONS OF THE HYPOTHALAMUS AND ASSOCIATED LIMBIC STRUCTURES

Aside from the vegetative and endocrine functions of the hypothalamus, experiments have shown that stimulation of or lesions in the hypothalamus often have profound effects on emotional behavior of animals or human beings.

In animals, some of the behavorial effects of stimulation are the following:

1. Stimulation in the *lateral hypothalamus* not only causes thirst and eating as discussed above in relation to the vegetative functions but also increases the general level of activity of the animal, sometimes leading to overt rage and fighting as will be discussed subsequently.

2. Stimulation in the *ventromedial nuclei* and

surrounding areas mainly causes effects opposite to those caused by lateral hypothalamic stimulation—that is, a sense of *satiety, decreased eating* and *tranquility.*

3. Stimulation of a *thin zone of periventricular nuclei,* located immediately adjacent to the third ventricle (or also stimulation of the central gray area of the mesencephalon that is continuous with this portion of the hypothalamus) often leads to *fear* and *punishment reactions*

4 *Sexual drive* can be stimulated from several areas of the hypothalamus, especially the most anterior and most posterior portions of the hypothalamus.

Lesions in the hypothalamus, in general, cause the opposite effects. For instance:

1. Bilateral lesions in the lateral hypothalamus will decrease drinking and eating almost to zero, often leading to lethal starvation, as was discussed above in relation to the vegetative functions of the lateral hypothalamus But these lesions cause extreme *passivity* of the animal as well, with loss of most of its overt activities.

2. Bilateral lesions of the ventromedial areas of the hypothalamus cause exactly opposite effects. excessive drinking and eating, as well as hyperactivity and often continuous savagery along with frequent bouts of extreme rage on the slightest provocation.

Lesions or stimulation in other regions of the limbic system, especially the amygdala, the septal area, and areas in the mesencephalon, often also cause effects similar to those elicited directly from the hypothalamus. We will discuss some of these in more detail later

The Reward and Punishment Function of the Limbic System

From the preceding discussion it is already clear that the hypothalamus and other limbic structures are particularly concerned with the affective nature of sensory sensations—that is, whether the sensations are *pleasant* or *unpleasant.* These affective qualities are also called *reward* and *punishment* or *satisfaction* and *aversion* Electrical stimulation of certain regions pleases or satisfies the animal, whereas electrical stimulation of other regions causes terror, pain, fear, defense, escape reactions, and all the other elements of punishment Obviously, these two oppositely responding systems greatly affect the behavior of the animal.

Reward Centers. Figure 56–4 illustrates a technique that has been used for localizing the specific reward and punishment areas of the brain In this figure a lever is placed at the side of the cage and is arranged so that depressing the lever makes electrical contact with a stimulator. Electrodes are placed successively at different areas in the brain so that the animal can stimulate the area by pressing the lever. If stimulating the rticular area gives the animal o sense of reward, then it will press the lever again and again, some times as much as 10,000 times per hour. Further more, when offered the choice of eating some de lectable food as opposed to the opportunity t stimulate the reward center, it often chooses th electrical stimulation.

By using this procedure, the major reward cen ters have been found to be located *along the cour of the medial forebrain bundle,* especially in th *lateral* and *ventromedial nuclei of the hypothal mus* It is stronge that the loterol nuclei should b included among the reward oreas—indeed it is on of the most potent reward oreas of all—becau even stronger stimuli in this orea con cause rag But this is true in many areas, with weaker stimu giving a sense of reward and stronger ones a sen of punishment

Less potent reward centers, which are perhap secondary to the major ones in the hypothalamu are found in the septum, the amygdala, certai areas of the thalamus and basal ganglia, an finally extending downward into the basal tegmen tum of the mesencephalon.

Punishment Centers. The apparatus illu trated in Figure 56–4 con also be connected so th pressing the lever turns off rather than turns o an electrical stimulus In this case, the anim will not turn the stimulus off when the electro is in one of the reward areas, but when it is certain other areas it immediately learns to tu it off Stimulation in these areas causes the anim to show all the signs of displeasure, fear, terr and punishment Furthermore, prolonged stim lation for 24 hours or more can cause the anim to become severely sick and actually lead to deat

By means of this technique, the most pote areas for punishment and escape tendencies ha been found in the *central gray area surroundi the aqueduct of Sylvius in the mesencephalon* a

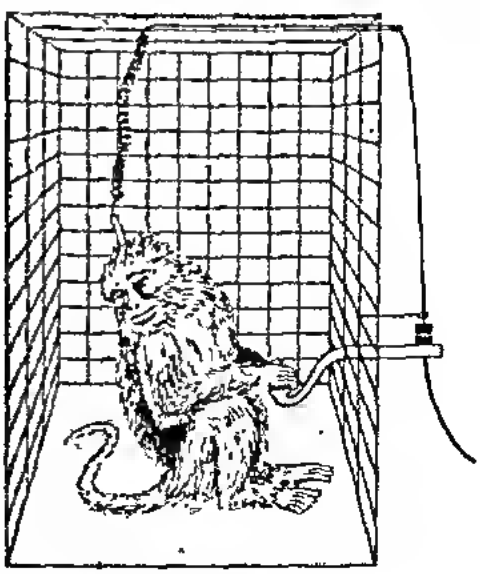


Figure 56–4. Technique for localizing reward and punishment centers in the brain of a monkey

extending upward into the *periventricular zones of the hypothalamus and thalamus*. Also, less potent punishment areas are found in the *amygdala* and the *hippocampus*.

It is particularly interesting that stimulation in the punishment centers can frequently inhibit the reward and pleasure centers completely, illustrating that punishment and fear can take precedence over pleasure and reward.

Importance of Reward and Punishment in Behavior. Almost everything that we do is related in some way to reward and punishment. If we are doing something that is rewarding, we continue to do it; if it is punishing, we cease to do it. Therefore, the reward and punishment centers undoubtedly constitute one of the most important of all the controllers of our bodily activities, our drives, and so forth.

Importance of Reward and Punishment in Learning and Memory—Habituation or Reinforcement. Animal experiments have shown that a sensory experience causing neither reward nor punishment is remembered hardly at all. Electrical recordings show that new and novel sensory stimuli always excite the cerebral cortex. But repetition of the stimulus over and over leads to almost complete extinction of the cortical response if the sensory experience does not elicit either a sense of reward or punishment. Thus, the animal becomes *habituated* to the sensory stimulus, and thereafter ignores it.

However, if the stimulus causes either reward or punishment rather than indifference, the cortical response becomes progressively more and more intense with repeated stimulation instead of fading away, and the response is said to be *reinforced*. Thus, an animal builds up strong memory traces for sensations that are either rewarding or punishing but, on the other hand, develops complete habituation to indifferent sensory stimuli. Therefore, it is evident that the reward and punishment centers of the limbic system have much to do with selecting the information that we learn

Effect of Tranquilizers on the Reward and Punishment Centers. Administration of a tranquilizer, such as chlorpromazine, inhibits both the reward and punishment centers, thereby greatly decreasing the affective reactivity of the animal. Therefore, it is presumed that tranquilizers function in psychotic states by suppressing many of the important behavioral areas of the hypothalamus and its associated regions of the brain, a subject that we will discuss more fully later.

Rage

An emotional pattern that involves the hypothalamus and has been well characterized is the *rage pattern*. This can be described as follows:

Strong stimulation of the punishment centers of the brain, especially the *periventricular zone of the hypothalamus* or in the *lateral hypothalamus*, causes the animal to (1) develop a defense posture, (2) extend its claws, (3) lift its tail, (4) hiss, (5) spit, (6) growl, and (7) develop piloerection, wide-open eyes, and dilated pupils. Furthermore, even the slightest provocation causes an immediate savage attack. This is approximately the behavior that one would expect from an animal being severely punished, and it is a pattern of behavior that is called *rage*.

Stimulation of the more rostral areas of the punishment areas—in the midline preoptic areas—causes mainly fear and anxiety, associated with a tendency for the animal to run away.

In the normal animal the rage phenomenon is held in check mainly by counterbalancing activity of the ventromedial nuclei of the hypothalamus. In addition, the hippocampus, the amygdala, and the anterior portions of the limbic cortex, especially the limbic cortex of the anterior cingulate gyrus and the subcallosal gyrus, help suppress the rage phenomenon. Conversely, if these portions of the limbic system are damaged or destroyed, the animal (also the human being) becomes far more susceptible to bouts of rage.

Placidity and Tameness. Exactly the opposite emotional behavioral patterns occur when the reward centers are stimulated: placidity and tameness.

Function of the Hypothalamus and Other Limbic Areas in Sleep, Wakefulness, Alertness, and Excitement

Stimulation of regions of the hypothalamus dorsal to the mammillary bodies greatly excites the reticular activating system and therefore causes wakefulness, alertness, and excitement. In addition, the sympathetic nervous system becomes excited in general, increasing the arterial pressure, causing pupillary dilatation, and enhancing other activities associated with sympathetic activity.

On the other hand, stimulation of some areas in the septum, in the anterior hypothalamus, or in isolated points of the thalamic portions of the reticular activating system often inhibits the mesencephalic portion of the reticular activating system, causing somnolence and sometimes actual sleep. Thus, the hypothalamus and other limbic structures indirectly contribute much to the control of the degree of excitement and alertness.

SPECIFIC FUNCTIONS OF OTHER PARTS OF THE LIMBIC SYSTEM

FUNCTIONS OF THE AMYGDALA

The amygdala is a complex of nuclei located immediately beneath the medial surface of the cerebral cortex in the anterior pole of each temporal lobe, and it has abundant bidirectional connections with the hypothalamus. In lower animals, this complex is concerned to a great extent with association of olfactory stimuli with

stimuli from other parts of the brain. Indeed, it is pointed out in Chapter 62 that one of the major divisions of the olfactory tract leads directly to a portion of the amygdala called the *corticomedial nuclei* that lie immediately beneath the cortex in the pyriform area of the temporal lobe. However, in the human being, another portion of the amygdala, the *basolateral nuclei*, has become much more highly developed than the olfactory portion and plays important roles in many behavioral activities not generally associated with olfactory stimuli.

The amygdala receives impulses from all portions of the limbic cortex, from the orbital surfaces of the frontal lobes, from the cingulate gyrus, from the parahippocampal gyrus, and from the neocortex of the temporal, parietal, and occipital lobes, especially from the auditory and visual association areas. Because of these multiple connections, the amygdala has been called the "window" through which the limbic system sees the place of the person in the world. In turn, the amygdala transmits signals (a) back into these same cortical areas, (b) into the hippocampus, (c) into the septum, (d) into the thalamus, and (e) especially into the hypothalamus.

Effects of Stimulating the Amygdala. In general, stimulation in the amygdala can cause almost all the same effects as those elicited by stimulation of the hypothalamus, plus still other effects. The effects that are mediated through the hypothalamus include (1) increases or decreases in arterial pressure, (2) increases or decreases in heart rate, (3) increases or decreases in gastrointestinal motility and secretion, (4) defecation and micturition, (5) pupillary dilatation or, rarely, constriction, (6) piloerection, (7) secretion of the various anterior pituitary hormones, including especially the gonadotropins and ACTH.

Aside from these effects mediated through the hypothalamus, amygdala stimulation can also cause different types of involuntary movement. These include (1) tonic movements, such as raising the head or bending the body, (2) circling movements, (3) occasionally clonic, rhythmic movements, and (4) different types of movements associated with olfaction and eating, such as licking, chewing, and swallowing.

In addition, stimulation of certain amygdaloid nuclei can, rarely, cause a pattern of rage, escape, punishment, and fear similar to the rage pattern elicited from the hypothalamus as described earlier. And stimulation of other nuclei can give reactions of reward and pleasure.

Finally, excitation of still other portions of the amygdala can cause sexual activities that include erection, copulatory movements, ejaculation, ovulation, uterine activity, and premature labor.

Effects of Bilateral Ablation of the Amygdala—The Klüver-Bucy Syndrome. When the anterior portions of both temporal lobes are destroyed in a monkey, this not only removes the temporal cortex but also the amygdalas that lie deep in these parts of the temporal lobes. This causes a combination of changes in behavior called the Klüver-Bucy syndrome, which includes (1) excessive tendency to examine objects orally, (2) loss of fear, (3) decreased aggressiveness, (4) tameness, (5) changes in dietary habits, even to the extent that a herbivorous animal frequently becomes carnivorous, (6) sometimes psychic blindness, and (7) often excessive sex drive. The characteristic picture is of an animal that is not afraid of anything, has extreme curiosity about everything, forgets very rapidly, has a tendency to place everything in its mouth and sometimes even tries to eat inedible objects, and, finally, often has a sex drive so strong that it attempts to copulate with immature animals, animals of the wrong sex, or animals of a different species.

Though similar lesions in human beings are rare, afflicted persons respond in a manner not too different from that of the monkey.

Overall Function of the Amygdala. The amygdala seems to be a behavioral awareness area that operates at a semiconscious level. It also seems to project into the limbic system one's present status in relation both to surroundings and thoughts. Based on this information, the amygdala is believed to help pattern the person's behavioral response so that it is appropriate for each occasion.

FUNCTIONS OF THE HIPPOCAMPUS

The hippocampus is an elongated structure composed of a modified type of cerebral cortex. In fact it is that portion of the temporal lobe cortex that folds inward to form the ventral surface of the inferior horn of the lateral ventricle. One end of the hippocampus abuts the amygdaloid nuclei, and it also fuses along one of its borders with the parahippocampal gyrus, which is the cortex of the ventromedial surface of the temporal lobe.

The hippocampus has numerous connections with most portions of the cerebral cortex as well as with the basic structures of the limbic system—the amygdala, the hypothalamus, the septum, and the mammillary bodies. Almost any type of sensory experience causes instantaneous activation of different parts of the hippocampus, and the hippocampus in turn distributes many outgoing signals to the hypothalamus and other parts of the limbic system, especially through the *fornix*, one of its major output pathways. Thus, the hippocampus, like the amygdala, is an additional channel through which incoming sensory signals can lead to appropriate limbic reactions, but perhaps for different purposes as we shall see later.

As in other limbic structures, stimulation of different areas in the hippocampus can cause almost any one of different behavioral patterns, such as rage, passivity, excess sex drive, or so forth.

Another feature of the hippocampus is that very weak electrical stimuli can cause local epileptic seizures that persist for many seconds after the stimulation is over, suggesting that the hippocampus can perhaps give off prolonged output signals even under normal functioning conditions. During the hippocampal seizures, the person experiences various psychomotor effects, including olfactory, visual, auditory, tactile, and other types of hallucinations that cannot be suppressed even though the person has not lost consciousness and knows these hallucinations to be unreal. Probably one of the reasons for this hyperexcitability of the hippocampus is that it is composed of a different type of cortex from that elsewhere in the cerebrum, having only three neuronal layers instead of the six layers found elsewhere.

Effect of Bilateral Removal of the Hippocampi—Inability to Learn. The hippocampi have been surgically removed bilaterally in a few human beings for the treatment of epilepsy. These persons can perform most previously learned activities satisfactorily. However, they can learn essentially nothing new. In fact, they cannot even learn the names or the faces of persons with whom they come in contact every day. Yet, they can remember for a moment or so what transpires during the course of their activities. Thus, they do have a type

f short-term primary memory even though their ability o establish new long-term secondary memories is either ompletely or almost completely abolished, which is the henomenon called *anterograde amnesia* that was dis- 1 in Chapter 54.

Destruction of the hippocampi also causes some deficit n previously learned memories (retrograde amnesia), a ittle more so for recent memories than for memories of he distant past.

Theoretical Function of the Hippocampus in .earning. The hippocampus originated as part of the lfactory cortex. In the very lowest animals it plays ssential roles in determining whether the animal will at a particular food, whether the smell of a particular bject suggests danger, whether the odor is sexually nviting, and in making other decisions that are of life ind death importance Thus, very early in the development of the brain, the hippocampus presumably became i critical decision-making neuronal mechanism, determining the importance and type of importance of the ncoming sensory signals. Presumably, as the remainder of the brain developed, the connections from the other sensory areas into the hippocampus have continued to utilize this decision-making capability.

Earlier in this chapter (and also in the previous chapter), it was pointed out that reward and punishment play a major role in determining the importance of information and especially whether or not the information will be stored in memory. A person rapidly becomes habituated to indifferent stimuli but learns assiduously any sensory experience that causes either pleasure or punishment. Yet, what is the mechanism by which this occurs? It has been suggested that the hippocampus provides the drive that causes translation of primary memory into secondary memory—that is, it transmits some type of signal to the long-term memory storage area directing that permanent storage shall take place.

Whatever the mechanism, without the hippocampi *consolidation* of long-term memories does not take place This is especially true for verbal information, perhaps because the temporal lobes, in which the hippocampi are located, are particularly concerned with verbal information.

FUNCTION OF THE LIMBIC CORTEX

Probably the most poorly understood portion of the entire limbic system is the ring of cerebral cortex called the *limbic cortex* that surrounds the subcortical limbic structures. This cortex functions as a transitional zone through which signals are transmitted from the remainder of the cortex into the limbic system. Therefore, it is presumed that the limbic cortex functions as a cerebral *association area for control of behavior.*

Stimulation of the different regions of the limbic cortex has failed to give any real idea of their functions. However, as is true of so many other portions of the limbic system, essentially all the behavioral patterns that have already been described can also be elicited by stimulation in different portions of the limbic cortex. Likewise, ablation of a few limbic cortical areas can cause persistent changes in an animal'e behavior, though ablation in most areas has little effect. Some of the areas that have been noted to cause specific changes include:

Ablation of the Temporal Cortex. When the anterior temporal cortex is ablated bilaterally, the amygdala is almost invariably damaged as well. This was discussed earlier, and it was pointed out that the Klüver-Bucy syndrome occurs. The animal develops consummatory behavior especially, investigates any and all objects, has intense sex drives toward inappropriate animals or even inanimate objects, and loses all fear—thus develops tameness as well.

Ablation of the Posterior Orbital Frontal Cortex. Bilateral removal of the posterior portion of the orbital frontal cortex often causes an animal to develop insomnia and an intense degree of motor restlessness, becoming unable to sit still but moving about continually.

Ablation of the Anterior Cingulate Gyri and Subcallosal Gyri. The anterior cingulate gyri and the subcallosal gyri are the portions of the limbic cortex that communicate between the prefrontal cerebral cortex and the subcortical limbic structures. Destruction of these gyri bilaterally releases the rage centers of the septum and hypothalamus from an inhibitory influence. Therefore, the animal can become vicious and much more subject to fits of rage than normally.

Summary. Until further information is available, it is perhaps best to state that the cortical regions of the limbic system occupy intermediate associative positions between the functions of the remainder of the cerebral cortex and the functions of the subcortical limbic structures for control of behavioral patterns. Thus, in the anterior temporal cortex one especially finds gustatory and olfactory associations In the parahippocampal gyri there is a tendency for complex auditory associations and also complex thought associations derived from Wernicke's area of the posterior temporal lobe. In the cingulate cortex, there is reason to believe that sensorimotor associations occur.

FUNCTION OF SPECIFIC CHEMICAL TRANSMITTER SYSTEMS FOR BEHAVIOR CONTROL

In previous chapters it has already been pointed out that several special collections of neurons in the brain stem radiate to multiple areas of the forebrain, each releasing a specific chemical transmitter substance. Three of these special chemical transmitter systems have received particular attention as behavior modifiers; these are the norepinephrine system, the serotonin system, and the dopamine system.

The Norepinephrine and Serotonin Systems and Their Relationship to Depression and Manic-Depressive Psychoses. In the previous chapter it was pointed out that large numbers of *norepinephrine-secreting neurons* are located in the reticular formation, especially in the *locus ceruleus,* and that these send fibers upward to most parts of the limbic system, the thalamus, and cerebral cortex. Also, many *serotonin-producing neurons* are located in the *midline raphe nuclei* of the lower pons and medulla and also project fibers to many areas of the limbic system and to some other areas of the brain as well.

In the past few years much evidence has accumulated to show that the *mental depression psychosis,* which afflicts about 8 million people in the United States at any one time, might be caused by diminished formation of either norepinephrine or serotonin or both. These patients experience symptoms of grief, unhappiness, despair, and misery. In addition, they lose their appetite and sex drive and also have severe insomnia. And

stimuli from other parts of the brain. Indeed, it is pointed out in Chapter 62 that one of the major divisions of the olfactory tract leads directly to a portion of the amygdala called the *corticomedial nuclei* that lie immediately beneath the cortex in the pyriform area of the temporal lobe. However, in the human being, another portion of the amygdala, the *basolateral nuclei*, has become much more highly developed than the olfactory portion and plays important roles in many behavioral activities not generally associated with olfactory stimuli

The amygdala receives impulses from all portions of the limbic cortex, from the orbital surfaces of the frontal lobes, from the cingulate gyrus, from the parahippocampal gyrus, and from the neocortex of the temporal, parietal, and occipital lobes, especially from the auditory and visual association areas Because of these multiple connections, the amygdala has been called the "window" through which the limbic system sees the place of the person in the world. In turn, the amygdala transmits signals (a) back into these same cortical areas, (b) into the hippocampus, (c) into the septum, (d) into the thalamus, and (e) especially into the hypothalamus

Effects of Stimulating the Amygdala. In general, stimulation in the amygdala can cause almost all the same effects as those elicited by stimulation of the hypothalamus, plus still other effects The effects that are mediated through the hypothalamus include (1) increases or decreases in arterial pressure, (2) increases or decreases in heart rate, (3) increases or decreases in gastrointestinal motility and secretion, (4) defecation and micturition, (5) pupillary dilatation or, rarely, constriction, (6) pilocrection, (7) secretion of the various anterior pituitary hormones, including especially the gonadotropins and ACTH,

Aside from these effects medisted through the hypothalamus, amygdala stimulation can also cause different types of involuntary movement These include (1) tonic movements, such as raising tha head or bending the body, (2) circling movements, (3) occasionally clonic, rhythmic movements, and (4) different types of movements associated with olfaction and eating, such as licking, chewing, and swallowing

In addition, stimulation of certain amygdaloid nuclei can, rarely, cause a pattern of raga, escape, punishment, and fear similar to the rage pattern elicited from the hypothalamus as described earlier And stimulation of other nuclei can give reactions of reward and pleasure.

Finally, excitation of still other portions of the amygdala can cause sexual activities that include erection, copulatory movements, ejaculation, ovulation, uterine activity, and premature labor

Effects of Bilateral Ablation of the Amygdala—The Klüver-Bucy Syndrome. When the anterior portions of both temporal lobes are destroyed in a monkey, this not only removes the temporal cortex but also the amygdalas that lie deep in these parts of the temporal lobes This causes a combination of changes in behavior called the Klüver-Bucy syndrome, which includes (1) excessive tendenry to examine objects orally, (2) loss of fear, (3) decreased aggressiveness, (4) tameness, (5) changes in dietary habits, even to the extent that a herbivorous animal frequently becomes carnivorous, (6) sometimes psychic blindness, and (7) often excessive sex drive The characteristic picture is of an animal that is not afraid of anything, has extreme ruriosity about everything, forgets very rapidly, has a tendency to place everything in its mouth and sometimes even tries to eat lid objects, and, finally, often has a sex drive so strong that it attempts to copulate with immature animals of the wrong sex, or animals of a ou species.

Though similar lesions in human beings are rar afflicted persons respond in a manner not too differen from that of the monkey.

Overall Function of the Amygdala. The amygdala seems to be a behavioral awareness area that operates at a semiconscious level It also seems to project into the limbic system one's present status in relation both to surroundings and thoughta Based on this information, the amygdala is believed to help pattern the person's behavioral response so that it is appropriate for each occasion.

FUNCTIONS OF THE HIPPOCAMPUS

The hippocampus is an elongated structure composed of a modified type of cerebral cortex. In fact it is that portion of the temporal lobe rortex that folds inward to form the ventral surface of the inferior horn of the lateral ventricle. One end of the hippocampus abuts the amygdaloid nuclei, and it also fuses along one of its borders with the parahippocampal gyrus, which is the cortex of the ventromedial surface of the temporal lobe.

The hippocampus has numerous connections with most portions of the cerebral cortex as well as with the basic structures of the limbic system—the amygdala, the hypothalamus, the septum, and the mammillary bodies Almost any type of sensory experience causes instantaneous activation of different parts of the hippocampus, and the hippocampus in turn distributes many outgoing signals to the hypothalamus and other parts of the limbic system, especially through the *fornix* one of its major output pathways Thus, tha hippocampus, like the amygdala, is an additional channel through which incoming sensory signals can lead to appropriate limbic reactions, but perhaps for different purposes as we shall see later

As in other limbic structures, stimulation of different areas in the hippocampus can cause almost any one of different behavioral patterns, such as rage, passivity excess sex drive, or so forth

Another feature of the hippocampus is that very weak electrical stimuli can cause local epileptic seizures that persist for many seconds after the stimulation is over, suggesting that the hippocampus can perhaps give off prolonged output signals even under normal functioning conditions During the hippocampal seizures, the person experiences various psychomotor effects, including olfactory, visual, auditory, tactile, and other types of hallucinations that cannot be suppressed even though the person has not lost consciousness and knows these hallucinations to be unreal. Probably one of the reasons for this hyperexcitability of the hippocampus is that it is composed of a different type of cortex from that elsewhere in the cerebrum, having only three neuronal layers instead of the six layers found elsewhere

Effect of Bilateral Removal of the Hippocampi—Inability to Learn. The hippocampi have been surgically removed bilaterally in a few human beings for the treatment of epilepsy. These persons can perform most previously learned activities satisfactorily However, they can learn essentially nothing new In fact, they cannot even learn the names or the faces of persons with whom they come in contact every day. Yet, they can remember for a moment or so what transpires during the course of their activities. Thus, they do have a type

short-term primary memory even though their ability establish new long-term secondary memories is either mpletely or almost completely abolished, which is the henomenon called *anterogrode amnesia* that was dis-ussed in Chapter 54.

Destruction of the hippocampi also causes some deficit previously learned memories (retrograde amnesia), a ttle more so for recent memories than for memories of e distant past.

Theoretical Function of the Hippocampus in earning. The hippocampus originated as part of the lfactory cortex. In the very lowest animals it plays ssential roles in determining whether the animal will at a particular food, whether the smell of a particular bject suggests danger, whether the odor is sexually viting, and in making other decisions that are of life nd death importance. Thus, very early in the develop-ent of the brain, the hippocampus presumably became critical decision-making neuronal mechanism, deter-ining the importance and type of importance of the coming sensory signals. Presumably, as the remainder f the brain developed, the connections from the other ensory areas into the hippocampus have continued to tilize this decision-making capability.

Earlier in this chapter (and also in the previous hapter), it was pointed out that reward and punishment lay a major role in determining the importance of nformation and especially whether or not the informa-ion will be stored in memory. A person rapidly becomes abituated to indifferent stimuli but learns assiduously ny sensory experience that causes either pleasure or unishment. Yet, what is the mechanism by which this ccurs? It has been suggested that the hippocampus rovides the drive that causes translation of primary nemory into secondary memory—that is, it transmits ome type of signal to the long-term memory storage rea directing that permanent storage shall take place

Whatever tha mechanism, without the hippocampi *onsolidation* of long-term memories does not take place This is especially true for verbal information, perhaps ecause the temporal lobes, in which the hippocampi re located, are particularly concerned with verbal in-ormation.

UNCTION OF THE LIMBIC CORTEX

Probably the most poorly understood portion of the entire limbic system is the ring of cerebral cortex called he *limbic cortex* that surrounds the subcortical limbic structures. This cortex functions as a transitional zone hrough which signals are transmitted from the remain-der of the cortex into the limbic system. Therefore, it is presumed that the limbic cortex functions as a cerebral *association area for control of behavior.*

Stimulation of the different regions of the limbic cortex has failed to give any real idea of their functions. However, as is true of so many other portions of the limbic system, essentially all the behavioral patterns that have already been described can also be elicited by stimulation in different portions of the limbic cortex. Likewise, ablation of a few limbic cortical areas can cause persistent changes in an animal's behavior, though ablation in most areas has little effect. Some of the areas that have been noted to cause specific changes include:

Ablation of the Temporal Cortex. When the ante-rior temporal cortex is ablated bilaterally, the amygdala is almost invariably damaged as well. This was dis-cussed earlier, and it was pointed out that the Klüver-Bucy syndrome occurs. The animal develops consum-matory behavior especially, investigates any and all objects, has intense sex drives toward inappropriate animals or even inanimate objects, and loses all fear—thus develops tameness as well.

Ablation of the Posterior Orbital Frontal Cortex. Bilateral removal of the posterior portion of the orbital frontal cortex often causes an animal to develop insom-nia and an intense degree of motor restlessness, becom-ing unable to sit still but moving about continually.

Ablation of the Anterior Cingulate Gyri and Sub-callosal Gyri. The anterior cingulate gyri and the subcallosal gyri are the portions of the limbic cortex that communicate between the prefrontal cerebral cor-tex and the subcortical limbic structures. Destruction of these gyri bilaterally releases the rage centers of the septum and hypothalamus from an inhibitory influence. Therefore, the animal can become vicious and much more subject to fits of rage than normally.

Summary. Until further information is available, it is perhaps best to state that the cortical regions of the limbic system occupy intermediate associative positions between the functions of the remainder of the cerebral cortex and the functions of the subcortical limbic struc-tures for control of behavioral patterns. Thus, in the anterior temporal cortex one especially finds gustatory and olfactory associations In tha parahippocampal gyri there is a tendency for complex auditory associations and also complex thought associations derived from Wernicke's area of the posterior temporal lobe. In the cingulate cortex, there is reason to beliave that senso-rimotor associations occur

FUNCTION OF SPECIFIC CHEMICAL TRANSMITTER SYSTEMS FOR BEHAVIOR CONTROL

In previous chapters it has already been pointed out that several special collections of neurons in the brain stem radiate to multiple areas of the forebrain, each releasing a specific chemical transmitter substance. Three of these special chemical transmitter systems have received particular attention as behavior modifiers; these are the norepinephrine system, the serotonin sys-tem, and the dopamine system.

The Norepinephrine and Serotonin Systems and Their Relationship to Depression and Manic-De-pressive Psychoses. In the previous chapter it was pointed out that large numbers of *norepinephrine-se-creting neurons* are located in the reticular formation, especially in the *locus ceruleus,* and that these send fibers upward to most parts of the limbic system, the thalamus, and cerebral cortex. Also, many *serotonin-producing neurons* are located in the *midline raphe nuclei* of the lower pons and medulla and also project fibers to many areas of the limbic system and to some other areas of the brain as well.

In the past few years much evidence has accumulated to show that the *mental depression psychosis,* which afflicts about 8 million people in the United States at any one time, might be caused by diminished formation of either norepinephrine or serotonin or both. These patients experience symptoms of grief, unhappiness, despair, and misery. In addition, they lose their appetite and sex drive and also have severe insomnia. And

associated with all these is a state of psychomotor agitation despite the depression

A principal reason for believing that depression is caused by diminished activity of the norepinephrine and serotonin systems is that drugs that block the secretion of norepinephrine and serotonin, such as the drug reserpine, frequently cause depression. Conversely, about 70 per cent of depressive patients can be treated very effectively with one of two types of drugs that increase norepinephrine and serotonin at the nerve endings (1) *monoamine oxidase inhibitors* that block destruction of norepinephrine and serotonin once they are formed, and (2) *tricyclic antidepressants* that block reuptake of norepinephrine and serotonin by the nerve endings so that these transmitters remain active for longer periods of time after secretion.

(Mental depression can also be treated effectively by electroconvulsive therapy—commonly called "shock therapy" In this therapy an electric shock is used to cause a generalized seizure similar to that of an epileptic attack. This causes considerable retrograde memory loss that recovers in about two to three weeks Mental depression is one of the very few mental illnesses in which such shock therapy is of real value However, the reason for its value is still to be determined)

Some patients with mental depression alternate between depression and mania, which is called the *manic-depressive psychosis*, and a few persons exhibit only mania without the depressive episodes. Drugs that block the formation or action of norepinephrine and serotonin, such as lithium compounds, can be effective in treating tha manic condition

Therefore, it is presumed that the norepinephrine and serotonin systems normally function to provide motor drive to the limbic system to increase a person's sense of well being, to create happiness, contentment, good appetite, appropriate sex drive, and psychomotor balance, though too much of a good thing can cause mania

The Dopamine System and Its Relationship to Schizophrenia. There are many reasons to believe that schizophrenia might be caused by excess secretion of dopamine in the brain In Chapter 52 we discussed the dopaminergic neurons that project from the substantia nigra to the basal ganglia and exert an important inhibitory restraint on basal ganglial activity In addition to the dopaminergic neurons in the substantia nigra, still other dopaminergic neurons are located in the ventral tegmentum of the mesencephalon medial and superior to the substantia nigra These neurons give rise to the so-called *mesolimbic dopaminergic system* that projects nerve fibers mainly into the medial and anterior portions of the limbic system, especially into the amygdala, the anterior caudate nucleus, and the anterior cingulate gyrus of the cortex, all of which are powerful behavioral control centers

Some of the reasons for believing the mesolimbic dopaminergic system to be related to schizophrenia are the following: When Parkinson's disease patients are treated with L-dopa, which releases dopamine in the brain, the parkinsonian patient sometimes develops schizophrenic symptoms, indicating that excess dopaminergic activity can cause dissociation of a person's drives and thought patterns However, an even more compelling reason for believing that schizophrenia might be caused by excess production of dopamine is that those drugs that are effective in treating schizophrenia, such as chlorpromazine and haloperidol, all [illegible] secretion of dopamine by the dopaminergic nerve endings or decrease the effect of dopamine o subsequent neurons

Almost certainly there are other factors in sc phrenia besides excess secretion of dopamine, but, n theless, the symptoms of schizophrenia are believ be similar to the behavioral effects of excessive mine as follows· (1) a sense of persecution from ou sources, (2) hearing of voices, (3) incoherent speec dissociation of ideas and abnormal sequences of tho and (5) sometimes abnormal postures or rigidities

The Enkephalin-Endorphin System. Very lit yet known about the behavioral functions of the phalin-endorphin system. However, the probable r this system for suppression of pain was discuss Chapter 50 It presumably has many other activiti well, because it is secreted in many areas of the stem and thalamus

Effects of the General Hormones on Beha Several of the general hormones have specific effe behavior Some of these are:

Thyroid hormone increases the overall *metaboli the body* and increases overall activity of the ne system as well.

Both *testosterone* in the male and *estrogen* i female increase *libido*. Also, injection of a minute tity of estrogen into the anterior hypothalamus of lower animals will cause estrous behavior (heat) sex drive of the human female seems to be incr near the time of ovulation when estrogen secret high

A small, 6 or 7 amino acid *fragment of ACTH*, injected into the brain of an animal, can cause *i fear*.

Injection of *norepinephrine*, or excess secretion o epinephrine or epinephrine by the adrenal med increases the overall activity of tha brain, wh essentially the same effect as stimulating the reti activating system

Thus, it is becoming clear that various chemica transmitter systems play important roles in determ an animal's behavior. Furthermore, abnormalit some of the brain's chemical transmitter systems to be the bases of some of the common psychoses

PSYCHOSOMATIC EFFECTS OF THE BEHAVIORAL SYSTEM

We are all familiar with the fact that abnormal function in the central nervous system can frequently lead to serious dysfunction of the different somatic organs of the body. This is also true in experimental animals Indeed, prolonged electrical stimulation in the punishment regions of the brain can actually lead to severe sickness of the animal, culminating in death within 24 to 48 hours. We need, therefore, to discuss briefly the mechanisms by which stimulatory effects in the brain can affect the peripheral organs. Ordinarily, this occurs through three routes: (1) through the motor nerves to the skeletal muscles throughout the body, (2) through the autonomic nerves to the different internal organs of the body, and (3) through the hormones secreted by the pituitary gland in response to nervous activity in the hypothalamus, as will be explained in Chapter 74

Psychosomatic Disorders Transmitted Through the Skeletal Nervous System. In neurotic and psy-

hotic states such as anxiety, tension, and mania, generalized overactivity of both the muscles and sympathetic system often occurs throughout the body. This in urn results in intense feedback from the muscle proioceptors to the reticular activating system, and the orepinephrine and epinephrine circulating in the blood as a result of the sympathetic activity also excite the eticular activating system, all of which undoubtedly elp maintain an extreme degree of wakefulness and lertness that characterizes these emotional states Unortunately, though, the wakefulness prevents adequate leep and also leads to progressive bodily fatigue and nental dissociations

Transmission of Psychosomatic Effects Through he Autonomic Nervous System. Many psychosomatic abnormalities result from hyperactivity of either the sympathetic or parasympathetic system In general, hyperactivity of the sympathetic system occurs in many areas of the body at the same time rather than in focal areas, and the usual effects are (1) increased heart rate—sometimes with palpitation of the heart, (2) increased arterial pressure, (3) constipation, and (4) increased metabolic rate. On the other hand, parasympathetic signals are likely to be much more focal. For instance, stimulation or inhibition of specific areas in the dorsal motor nuclei of the vagus nerves can cause more or less specifically (1) increased or decreased heart rate or palpitation of the heart, (2) esophageal spasm, (3) increased peristalsis in the upper gastrointestinal tract, or (4) increased hyperacidity of the stomach with resultant development of peptic ulcer Stimulation of the sacral region of the parasympathetic system, on the other hand, is likely to cause extreme colonic glandular secretion and peristalsis with resulting diarrhea

One can readily see, then, that emotional patterns controlling the sympathetic and parasympathetic centers of the hypothalamus and lower brain stem can cause wide varieties of peripheral psychosomatic effects

Psychosomatic Effects Transmitted Through the Anterior Pituitary Gland. Electrical stimulation of the midbasal hypothalamus increases the secretion of orticotropin (ACTH) by the anterior pituitary gland (as explained earlier in the chapter) and therefore indirectly increases the output of adrenocortical hormones One of the effects of this is a gradual increase in stomach hyperacidity because of the effect of glucocorticoids on stomach secretion. Over a prolonged period of time this obviously could lead to peptic ulcer, which is a well-known effect of hypersecretion by the adrenal cortex. Likewise, activity in the preoptic area of the hypothalamus increases the pituitary secretion of thyrotropin, which in turn increases the output of thyroxine and leads to an elevated basal metabolic rate. It is well known that different types of emotional disturbances can lead to thyrotoxicosis (as will be explained in Chapter 76), this presumably resulting from overactivity in the hypothalamus.

From these examples, therefore, it is evident that many types of psychosomatic diseases of the body can be caused by abnormal control of anterior pituitary secretion. These will be discussed more extensively in the section on endocrinology later in this text.

REFERENCES

Baldessarini, R. J, and Tarsy, D Dopamine and the pathophysiology of dyskinesias induced by antipsychotic drugs. *Annu Rev Neurosci*, 3 23, 1980

Berger, P A, *et al* Behavioral pharmacology of the endorphins. *Annu Rev Med* 33 397, 1982

Burchfield, S R. (ed.) Stress Physiological and Psychological Interactions Washington, D C, Hemisphere Publishing Corp, 1985

Buzsaki G Feed forward inhibition in the hippocampal formation. *Prog Neurobiol*, 22 131, 1984

Cohen, D H, and Randall, D C Classical conditioning of cardiovascular responses *Annu Rev Physiol* 46 187, 1984

Colman, C W, and McGaugh, J L Behavioral Neuroscience New York Academic Press, 1979

de Wied D and Jolles, J Neuropeptides derived from pro-opiocortin Behavioral, physiological, and neurochemical effects *Physiol Rev*, 62 976, 1982

Doris P A Vasopressin and central integrative processes *Neuroendocrinology* 38 75 1984

Engel B T and Schneiderman, N Operant conditioning and the modulation of cardiovascular function Annu. Rev Physiol., 46 199, 1984

Fink M Convulsive and drug therapies of depression Annu Rev Med, 32 405 1981

Foote S L, et al Nucleus locus ceruleus. New evidence of anatomical and physiological specificity *Physiol Rev*, 63 844, 1983

Fuller R W Pharmacology of brain epinephrine neurons. *Annu Rev Pharmacol Toxicol* 22 31, 1982

Ganong W F The brain renin-angiotensin system *Annu Rev Physiol*, 46 17 1984

Gwens, J R The Hypothalamus in Health and Disease Chicago, Year Book Medical Publishers, 1984

Guillemin R Neuroendocrine interrelations *In* Bondy, P K and Rosenberg L E (eds.) Metabolic Control and Disease, 8th Ed Philadelphia, W B Saunders Co 1980 p 1155

Iversen L L Nonopioid neuropeptides in mammalian CNS Annu Rev Pharmacol Toxicol 23 1 1983

Kelly D Anxiety and Emotions Physiological Basis and Treatment. Springfield Ill., Charles C Thomas, 1979

Klimov P K Behavior of the organs of the digestive system *Neurosci Behav Physiol* 14 333 1984

Lechtenberg R Psychiatrist's Guide to Diseases of the Nervous System. New York John Wiley & Sons 1982

Malick, J B (ed.) Anxiolytics Neurochemical, Behavioral and Clinical Perspectives New York Raven Press 1983

McFadden, D (ed.) Neural Mechanisms in Behavior New York, Springer Verlag, 1980

Russell, R W Cholinergic system in behavior The search for mechanisms of action Annu Rev Pharmacol Toxicol 22 435 1982

Saddle, D Orienting and Habituation Perspectives in Human Research. New York, John Wiley & Sons, 1982

Skodol, A E, and Spitzer, R. L. The development of reliable diagnostic criteria in psychiatry *Annu Rev Med* 33 317 1982

Smith, O A and DeVito, J L Central neural integration for the control of autonomic responses associated with emotion *Annu Rev Neurosci*, 7 43 1984

Snyder, S Biological Aspects of Mental Disorders New York, Oxford University Press, 1980

Stephenson, R B Modification of reflex regulation of blood pressure by behavior *Annu Rev Physiol*, 46 133 1984

Swanson, L W, and Sawchenko, P E Hypothalamic integration Organization of the paraventricular and supraoptic nuclei *Annu Rev Neurosci*, 6 269, 1983

Tucek, S Regulation of acetylcholine synthesis in the brain *J Neurochem*, 44 11, 1985

Usdin, E Stress The Role of Catecholamines and Other Neurotransmitters New York, Gordon Press Publishers, 1984

Verrier, R L, and Lown, B. Behavioral stress and cardiac arrhythmias. *Annu Rev Physiol*, 46 155, 1984

57

The Autonomic Nervous System; The Adrenal Medulla

The portion of the nervous system that controls the visceral functions of the body is called the *autonomic nervous system.* This system helps control arterial pressure, gastrointestinal motility and secretion, urinary bladder emptying, sweating, body temperature, and many other activities, some of which are controlled almost entirely and some only partially by the autonomic nervous system.

One of the most striking characteristics of the autonomic nervous system is the rapidity and intensity with which it can change visceral functions. For instance, within 3 to 5 seconds it can increase the heart rate to two times normal, and the arterial pressure can be doubled within as little as 10 to 15 seconds; or, at the other extreme, the arterial pressure can be decreased low enough within 4 to 5 seconds to cause a faint Sweating can begin within seconds, and the bladder may empty involuntarily, also within seconds Thus, the nervous system, acting through the autonomic nerve fibers, can exert rapid and effective control of many if not most of the internal functions of the body It is these extremely rapid changes that are measured by the lie detector polygraph, reflecting the innermost feelings of a person.

GENERAL ORGANIZATION OF THE AUTONOMIC NERVOUS SYSTEM

The autonomic nervous system is activated mainly by centers located in the *spinal cord, brain stem,* and *hypothalamus.* Also, portions of the cerebral cortex and especially of the limbic system can transmit impulses to the lower centers and in this way influence autonomic control. Often the autonomic nervous system also operates by means of *visceral reflexes.* That is, sensory signals enter the centers of the autonomic ganglia, cord, brain stem, or hypothalamus, and these in turn transmit appropriate reflex responses back to the visceral organs to control their activities.

The autonomic signals are transmitted to the through two major subdivisions called the *sympathetic* and *parasympathetic systems,* th characteristics and functions of which follow.

PHYSIOLOGIC ANATOMY OF THE SYMPATHETIC NERVOUS SYSTEM

Figure 57–1 illustrates the general organization the sympathetic nervous system, showing one of the tw

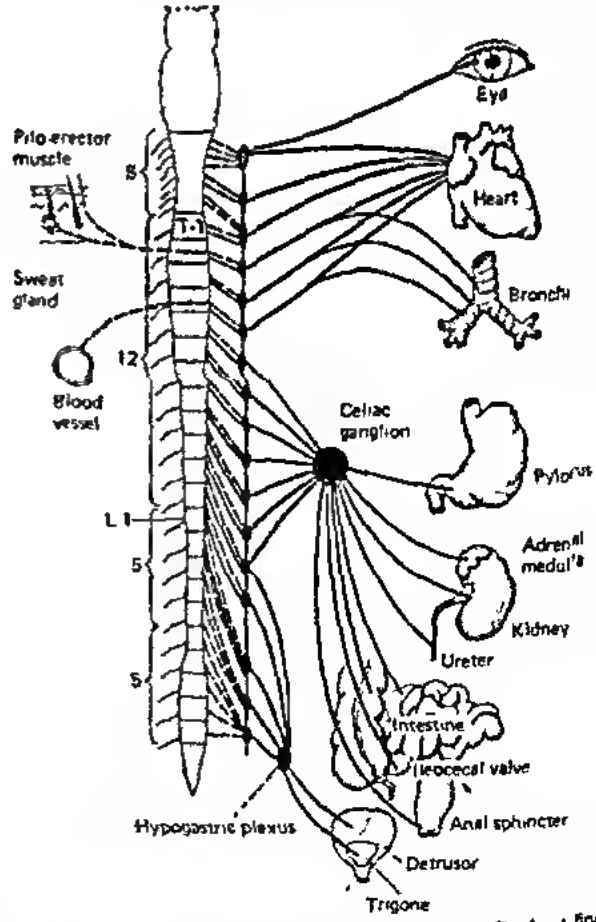


Figure 57–1. The sympathetic nervous system. Dashed lines represent postganglionic fibers in the gray rami leading into the spinal nerves for distribution to blood vessels, sweat glands, and pilloerector muscles

paravertebral sympathetic chains to the side of the spinal column and nerves extending to the different internal organs. The sympathetic nerves originate in the spinal cord between the segments T-1 and L-2 and pass from here first into the sympathetic chain, thence to the tissues and organs that are stimulated by the sympathetic nerves.

Preganglionic and Postganglionic Sympathetic Neurons. The sympathetic nerves are different from skeletal motor nerves in the following way Each motor pathway from the cord to a skeletal muscle is comprised of a single fiber. Each sympathetic pathway, on the other hand, is comprised of two fibers, a *preganglionic neuron* and a *postganglionic neuron*. The cell body of a preganglionic neuron lies in the *intermediolateral horn* of the spinal cord, and its fiber passes, as illustrated in Figure 57–2, through an *anterior root* of the cord into a *spinal nerve*.

Immediately after the spinal nerve leaves the spinal column, the preganglionic sympathetic fibers leave the nerve and pass through the *white ramus* into one of the *ganglia* of the *sympathetic chain*. Then the course of the fibers can be one of the following three: (1) It can synapse with postganglionic neurons in the ganglion that it enters. (2) It can pass upward or downward in the chain and synapse in one of the other ganglia of the chain Or (3) it can pass for variable distances through the chain and then through one of the nerves radiating outward from the chain, finally terminating in an *outlying sympathetic ganglion*.

The postganglionic neuron, therefore, can originate either in one of the sympathetic chain ganglia or in one of the outlying ganglia. From either of these two sources, the postganglionic fibers then travel to their destinations in the various organs.

Sympathetic Nerve Fibers in the Skeletal Nerves. Many of the postganglionic fibers pass back from the sympathetic chain into the spinal nerves through *gray rami* (see Fig. 57–2) at all levels of the cord These pathways are made up of type C fibers that extend to all parts of the body in the skeletal nerves They control the blood vessels, sweat glands, and piloerector muscles of the hairs Approximately 8 per cent of the fibers in the average skeletal nerve are sympathetic fibers, a fact that indicates their importance.

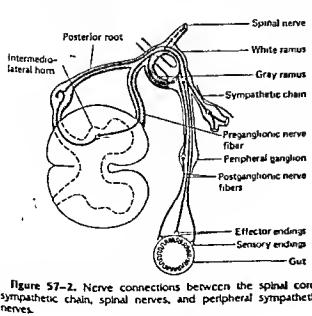


Figure 57–2. Nerve connections between the spinal cord, sympathetic chain, spinal nerves, and peripheral sympathetic nerves.

Segmental Distribution of Sympathetic Nerves. The sympathetic pathways originating in the different segments of the spinal cord are not necessarily distributed to the same part of the body as the spinal nerve fibers from the same segments. Instead, the *sympathetic fibers from T-1 generally pass up the sympathetic chain into the head; from T-2 into the neck; T-3, T-4, T-5, and T-6 into the thorax; T-7, T-8, T-9, T-10, and T-11 into the abdomen;* and *T-12, L-1, and L-2 into the legs*. This distribution is only approximate and overlaps greatly.

The distribution of sympathetic nerves to each organ is determined partly by the position in the embryo at which the organ originates For instance, the heart receives many sympathetic nerves from the neck portion of the sympathetic chain because the heart originates in the neck of the embryo. Likewise, the abdominal organs receive their sympathetic innervation from the lower thoracic segments because the primitive gut originates in this area.

Special Nature of the Sympathetic Nerve Endings in the Adrenal Medullae. Preganglionic sympathetic nerve fibers pass, without synapsing, all the way from the intermediolateral horn cells of the spinal cord, through the sympathetic chains, through the splanchnic nerves, and finally into the adrenal medullae. There they end directly on special cells that secrete epinephrine and norepinephrine directly into the circulatory blood. These secretory cells embryologically are derived from nervous tissue and are analogous to postganglionic neurons; indeed, they even have rudimentary nerve fibers, and it is these fibers that secrete the hormones.

PHYSIOLOGIC ANATOMY OF THE PARASYMPATHETIC NERVOUS SYSTEM

The parasympathetic nervous system is illustrated in Figure 57–3, showing that parasympathetic fibers leave the central nervous system through several of the cranial nerves, the second and third sacral spinal nerves, and occasionally the first and fourth sacral nerves. About 75 per cent of all parasympathetic nerve fibers are in the vagus nerves, passing to the entire thoracic and abdominal regions of the body Therefore, a physiologist speaking of the parasympathetic nervous system often thinks mainly of the two vagus nerves. The vagus nerves supply parasympathetic nerves to the heart, the lungs, the esophagus, the stomach, the small intestine, the proximal half of the colon, the liver, the gallbladder, the pancreas, and the upper portions of the ureters.

Parasympathetic fibers in the *third nerve* flow to the pupillary sphincters and ciliary muscles of the eye. Fibers from the *seventh nerve* pass to the lacrimal, nasal, and submandibular glands, and fibers from the *ninth nerve* pass to the parotid gland.

The sacral parasympathetic fibers congregate in the form of the *nervi erigentes*, also called the *pelvic nerves*, which leave the sacral plexus on each side of the cord and distribute their peripheral fibers to the descending colon, rectum, bladder, and lower portions of the ureters. Also, this sacral group of parasympathetics supplies fibers to the external genitalia to cause various sexual reactions.

Preganglionic and Postganglionic Parasympathetic Neurons. The parasympathetic system, like the sympathetic, has both preganglionic and postganglionic

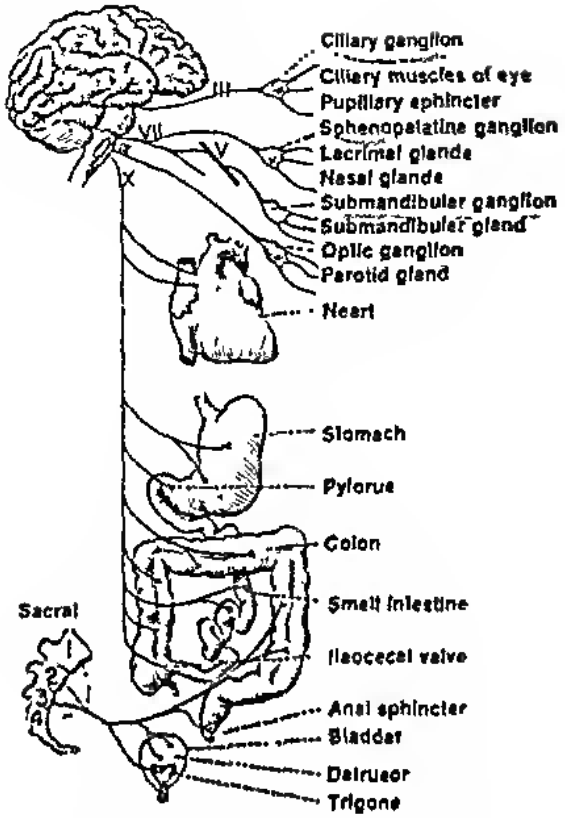


Figure 57–3. The parasympathetic nervous system.

neurons, but, except in the case of a few cranial parasympathetic nerves, the *preganglionic fibers* pass uninterrupted to the organ that is to be controlled. In the wall of the organ are located the *postganglionic neurons* of the parasympathetic system. The preganglionic fibers synapse with these; then short postganglionic fibers, 1 millimeter to several centimeters in length, leave the neurons to spread in the substance of the organ. This location of the parasympathetic postganglionic neurons in the visceral organ itself is quite different from the arrangement of the sympathetic ganglia, for the cell bodies of the sympathetic postganglionic neurons are almost always located in ganglia of the sympathetic chain or in various other discrete ganglia in the abdomen rather than in the excited organ itself.

BASIC CHARACTERISTICS OF SYMPATHETIC AND PARASYMPATHETIC FUNCTION

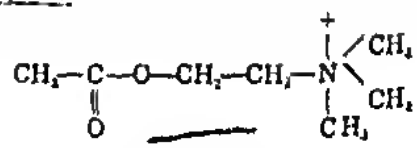
CHOLINERGIC AND ADRENERGIC FIBERS—SECRETION OF ACETYLCHOLINE OR NOREPINEPHRINE BY THE POSTGANGLIONIC NEURONS

The sympathetic and parasympathetic nerve endings secrete one of the two synaptic transmitter substances, *acetylcholine* or *norepinephrine*. Those fibers that secrete acetylcholine are said to be *cholinergic*. Those that secrete norepinephrine are said to be *adrenergic*, a term derived from adren- which is the British name for epinephrine.

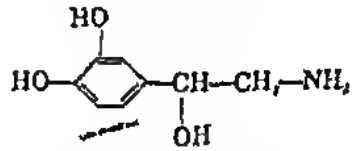
All *preganglionic neurons* are *cholinergic* in both the sympathetic and parasympathetic nervous systems. Therefore, acetylcholine or acetylcholine-like substances, when applied to the ganglia, will excite both sympathetic and parasympathetic postganglionic neurons.

The *postganglionic neurons of the parasympathetic system are* also *all cholinergic*. On the other hand, *most of the postganglionic sympathetic neurons are adrenergic*, though this is not entirely so because the sympathetic nerve fibers to the sweat glands, to the piloerector muscles, and to a few blood vessels are cholinergic.

Thus, in general, the terminal nerve endings of the parasympathetic system secrete acetylcholine, and most of the sympathetic nerve endings secrete norepinephrine. These hormones, in turn, act on the different organs to cause the respective parasympathetic and sympathetic effects. Therefore, these substances are often called respectively *parasympathetic* and *sympathetic transmitters*. These are the molecular structures of acetylcholine and norepinephrine.



Acetylcholine



Norepinephrine

Mechanism of Secretion of Acetylcholine and Norepinephrine by Autonomic Nerve Endings. Some of the autonomic nerve endings, especially those of the parasympathetic nerves, are similar to but much smaller in size than those of the skeletal neuromuscular junction. However, most of the sympathetic nerve fibers merely touch the effector cells of the organs that they innervate as they pass by; in some instances they terminate in connective tissue located adjacent to the cells that are to be stimulated. Where these filaments pass over or near the effector cells, they usually have bulbous enlargements called *varicosities*, and it is in these varicosities that the transmitter vesicles of acetylcholine and norepinephrine are found. Also in the varicosities are large numbers of mitochondria to supply the ATP required to energize acetylcholine or norepinephrine synthesis.

When an action potential spreads over the terminal fibers, the depolarization process increases the permeability of the fiber membrane to calcium ions, thus allowing these to diffuse in moderate numbers into the nerve terminals. There they interact with the vesicles adjacent to the membrane, causing them to fuse with the membrane and to empty their contents to the exterior. Thus, the transmitter substance is secreted.

Synthesis of Acetylcholine, Its Destruction After Secretion, and Duration of Action. Acetylcholine is

synthesized in the terminal endings of cholinergic nerve fibers. Most of this synthesis occurs in the axoplasm, and most of the acetylcholine is then transported to the interior of the vesicles. The basic chemical reaction of this synthesis is the following:

$$\text{Acetyl-CoA} + \text{Choline} \xrightarrow{\text{choline acetyl-transferase}} \text{Acetylcholine}$$

Once the acetylcholine has been secreted by the cholinergic nerve ending, most of it is split into acetate ion and choline by the enzyme *acetylcholinesterase* that is present in the synaptic area bound with the local collagen and glycosaminoglycans. Thus, this is the same mechanism of acetylcholine destruction that occurs at neuromuscular junctions of skeletal nerve fibers. The choline that is formed is in turn transported back into the terminal nerve ending where it is used again for synthesis of new acetylcholine. Though most of the acetylcholine is usually destroyed within a fraction of a second after its secretion, it sometimes persists for as long as several seconds, and a small amount also diffuses into the surrounding fluids. These fluids contain a different type of cholinesterase called *serum cholinesterase*, that destroys the remaining acetylcholine within another few seconds. Therefore, the action of acetylcholine released by cholinergic nerve fibers usually lasts for a few seconds at most.

Synthesis of Norepinephrine, Its Removal, and Duration of Action. Synthesis of norepinephrine begins in the axoplasm of the terminal nerve endings of adrenergic nerve fibers but is completed inside the vesicles. The basic steps are the following:

1. Tyrosine $\xrightarrow{\textit{hydroxylation}}$ DOPA
2. DOPA $\xrightarrow{\textit{decarboxylation}}$ Dopamine
3. Transport of dopamine into the vesicles
4. Dopamine $\xrightarrow{\textit{hydroxylation}}$ Norepinephrine

In the adrenal medulla this reaction goes still one step further to form epinephrine, as follows

5. Norepinephrine $\xrightarrow{\textit{methylation}}$ Epinephrine

Following secretion of norepinephrine by the terminal nerve endings, it is removed from the secretory site in three different ways: (1) re-uptake into the adrenergic nerve endings themselves by an active transport process—accounting for removal of 50 to 80 per cent of the secreted norepinephrine; (2) diffusion away from the nerve endings into the surrounding body fluids and thence into the blood—accounting for removal of most of the remainder of the norepinephrine; and (3) destruction by enzymes to a slight extent (one of these enzymes is *monoamine oxidase*, which is found in the nerve endings themselves, and another is *catechol-O-methyl transferase*, which is present diffusely in all tissues).

Ordinarily, the norepinephrine secreted directly in a tissue by adrenergic nerve endings remains active for only a few seconds, illustrating that its re-uptake and diffusion away from the tissue are rapid. However, the norepinephrine and epinephrine secreted into the blood by the adrenal medullae remain active until they diffuse into some tissue where they are destroyed by catechol-O-methyl transferase; this occurs mainly in the liver. Therefore, when secreted into the blood, both norepinephrine and epinephrine remain very active for 10 to 30 seconds, followed by decreasing activity thereafter for 1 to several minutes.

RECEPTORS OF THE EFFECTOR ORGANS

The acetylcholine, norepinephrine, and epinephrine transmitters secreted by the autonomic nervous system all stimulate the effector organs by first binding with *receptors* of the effector cells. The receptor in most instances is in the cell membrane and is a protein molecule. The usual mechanism for function of the receptor is that the transmitter first binds with the receptor and this causes a basic change in the structure of the protein molecule in the manner described in Chapters 9 and 10. Because the receptor is an integral part of the cell membrane, this structural change often opens or closes *ion channels*, thus altering the permeability of the cell membrane to various ions—for instance, to allow rapid influx or to diminish influx of sodium, chloride, or calcium ions into the cell or to alter the efflux of potassium ions out of the cell. These ionic changes then usually alter the membrane potential, sometimes causing increased negativity (hyperpolarization) or at other times causing decreased negativity (depolarization) to elicit action potentials (as occurs in some smooth muscle cells) and at other times causing electrotonic effects on the cells (as occurs in glandular cells) to produce the responses. The ions themselves have direct effects within the receptor cells, such as the effect of calcium ions to promote smooth muscle contraction.

Another way that the receptor can function, besides changing the membrane permeability, is to activate an enzyme in the cell membrane; this enzyme in turn promotes chemical reactions within the cell. For instance, epinephrine increases the activity of *adenyl cyclase* in some cell membranes, and this then causes the formation of cyclic AMP at the inner surfaces of the membranes; the cyclic AMP then initiates many intracellular activities.

The effect that occurs in each instance is determined by the nature of the receptors, as we shall see in the following paragraphs.

The Acetylcholine Receptors—"Muscarinic" and "Nicotinic" Receptors. Acetylcholine activates two different types of receptors. These are called *muscarinic* and *nicotinic* receptors. The reason for these names is that muscarine, a poison from toadstools, also activates the muscarinic receptors but will not activate the nicotinic receptors, whereas nicotine will activate only the other receptors; acetylcholine activates both of them.

The muscarinic receptors are found in all the effector cells stimulated by the postganglionic neurons of the parasympathetic nervous system, as

well as those stimulated by the postganglionic cholinergic neurons of the sympathetic system.

The nicotinic receptors are found in the synapses between the pre- and postganglionic neurons of both the sympathetic and parasympathetic systems and also in the membranes of skeletal muscle fibers at the neuromuscular junction (discussed in Chapter 12).

An understanding of the two different types of receptors is especially important because specific drugs are frequently used in the practice of medicine to stimulate or to block one or the other of the two types of receptors

The Adrenergic Receptors—"Alpha" and "Beta" Receptors. Research experiments using different drugs (called *sympathomimetic drugs*) that mimic the action of norepinephrine on sympathetic effector organs have shown that there are two major types of adrenergic receptors, called *alpha receptors* and *beta receptors*. (The beta receptors in turn are divided into *beta*$_1$ and *beta*$_2$ receptors because certain drugs affect some beta receptors but not all.)

Norepinephrine and epinephrine, both of which are secreted by the adrenal medulla, have somewhat different effects in exciting the alpha and beta receptors. Norepinephrine excites mainly alpha receptors but excites the beta receptors to a very slight extent as well On the other hand, epinephrine excites both types of receptors approximately equally. Therefore, the relative effects of norepinephrine and epinephrine on different effector organs is determined by the types of receptors in the organs Obviously, if they are all beta receptors, epinephrine will be the more effective excitant.

Table 57–1 gives the distribution of alpha and beta receptors in some of the organs and systems controlled by the sympathetics. Note that certain alpha functions are excitatory while others are inhibitory. Likewise, certain beta functions are excitatory and others are inhibitory. Therefore, alpha and beta receptors are not necessarily associated with excitation or inhibition but simply with the affinity of the hormone for the receptors in a given effector organ

A synthetic hormone chemically similar to epinephrine and norepinephrine, *isopropyl norepinephrine*, has an extremely strong action on beta receptors but essentially no action on alpha receptors. Later in the chapter we will discuss various drugs that can mimic the adrenergic actions of epinephrine and norepinephrine or that will block specifically the alpha or beta receptors.

Table 57–1 ADRENERGIC RECEPTORS AND FUNCTION

Alpha Receptor	Beta Receptor
Vasoconstriction	1 Vasodilatation (β_2)
Iris dilatation	2 Cardioacceleration (β_1)
Intestinal relaxation	3 Increased myocardial strength (β_1)
Intestinal sphincter contraction	4 Intestinal relaxation (β_2)
	5 Uterus relaxation (β_2)
Pilomotor contraction	6 Bronchodilatation (β_2)
Bladder sphincter contraction	7 Calorigenesis (β_2)
	8 Glycogenolysis (β_2)
	9 Lipolysis (β_1)
	10 Bladder relaxation (β_2)

EXCITATORY AND INHIBITORY ACTIONS OF SYMPATHETIC AND PARASYMPATHETIC STIMULATION

Table 57–2 lists the effects on different visceral functions of the body caused by stimulating the parasympathetic and sympathetic nerves. From this table it can be seen that *sympathetic stimulation causes excitatory effects in some organs but inhibitory effects in others. Likewise, parasympathetic stimulation causes excitation in some organs but inhibition in others.* Also, when sympathetic stimulation excites a particular organ, parasympathetic stimulation sometimes inhibits it, illustrating that the two systems occasionally act reciprocally to each other. However, most organs are dominantly controlled by one or the other of the two systems

There is no generalization one can use to explain whether sympathetic or parasympathetic stimulation will cause excitation or inhibition of a particular organ. Therefore, to understand sympathetic and parasympathetic function, one must learn the functions of these two nervous systems as listed in Table 57–2. Some of these functions need to be clarified in still greater detail as follows:

EFFECTS OF SYMPATHETIC AND PARASYMPATHETIC STIMULATION ON SPECIFIC ORGANS

The Eye. Two functions of the eye are controlled by the autonomic nervous system These are the pupillary opening and the focus of the lens. Sympathetic stimulation contracts the meridional *fibers of the iris* and, therefore, dilates the pupil, while parasympathetic stimulation contracts the *circular muscle of the iris* to constrict the pupil The parasympathetics that control the pupil are reflexly stimulated when excess light enters the eyes; this reflex reduces the pupillary opening and decreases the amount of light that strikes the retina On the other hand, the sympathetics become stimulated during periods of excitement and, therefore, increase the pupillary opening at these times

Focusing of the lens is controlled almost entirely by the parasympathetic nervous system The lens is normally held in a flattened state by tension of its radial ligaments. Parasympathetic excitation contracts the *ciliary muscle*, which releases this tension and allows the lens to become more convex. This causes the eye to focus on objects near at hand. The focusing mechanism is discussed in Chapters 58 and 60 in relation to function of the eyes

The Glands of the Body. The *nasal, lacrimal, salivary*, and many *gastrointestinal glands* are all strongly stimulated by the parasympathetic nervous system, resulting in copious quantities of secretion. The glands of the alimentary tract most strongly stimulated by the

Table 57–2 AUTONOMIC EFFECTS ON VARIOUS ORGANS OF THE BODY

	Organ	Effect of Sympathetic Stimulation	Effect of Parasympathetic Stimulation
1	Eye: Pupil	Dilated	Constricted
	Eye: Ciliary muscle	Slight relaxation	Constricted
2	Glands: Nasal, Lacrimal, Parotid, Submandibular, Gastric, Pancreatic	Vasoconstriction and slight secretion	Stimulation of copious (except pancreas) secretion (containing many enzymes for enzyme-secreting glands)
3	Sweat glands	Copious sweating (cholinergic)	None
4	Apocrine glands	Thick, odoriferous secretion	None
5	Heart: Muscle	Increased rate	Slowed rate
		Increased force of contraction	Decreased force of contraction (especially of atrium)
	Coronaries	Dilated (β_2), constricted (α)	Dilated
6	Lungs: Bronchi	Dilated	Constricted
	Blood vessels	Mildly constricted	? Dilated
7	Gut: Lumen	Decreased peristalsis and tone	Increased peristalsis and tone
	Sphincter	Increased tone (most times)	Relaxed (most times)
8	Liver	Glucose released	Slight glycogen synthesis
9	Gallbladder and bile ducts	Relaxed	Contracted
10	Kidney	Decreased output and renin secretion	None
11	Bladder: Detrusor	Relaxed (slight)	Excited
	Trigone	Excited	Relaxed
12	Penis	Ejaculation	Erection
13	Systemic arterioles: Abdominal	Constricted	None
	Muscle	Constricted (adrenergic α)	None
		Dilated (adrenergic β_2)	
		Dilated (cholinergic)	
	Skin	Constricted	None
14	Blood: Coagulation	Increased	None
	Glucose	Increased	None
15	Basal metabolism	Increased up to 100%	None
16	Adrenal medullary secretion	Increased	None
17	Mental activity	Increased	None
18	Piloerector muscles	Excited	None
19	Skeletal muscle	Increased glycogenolysis	None
		Increased strength	

parasympathetics are those of the upper tract, especially those of the mouth and stomach. The glands of the small and large intestines are controlled principally by local factors in the intestinal tract itself and not by the autonomic nerves.

Sympathetic stimulation has a slight direct effect on glandular cells in causing formation of a concentrated secretion. However, it also causes vasoconstriction of the blood vessels supplying the glands and in this way often reduces their rates of secretion.

The *sweat glands* secrete large quantities of sweat when the sympathetic nerves are stimulated, but no effect is caused by stimulating the parasympathetic nerves. However, the sympathetic fibers to most sweat glands are *cholinergic* (except for a few adrenergic fibers to palms of the hand and the soles of the feet), in contrast to most other sympathetic fibers, which are adrenergic. Furthermore, the sweat glands are stimulated primarily by centers in the (hypothalamus) that are usually considered to be parasympathetic centers. Therefore, sweating could be called a parasympathetic function.

The *apocrine glands* secrete a thick, odoriferous secretion as a result of sympathetic stimulation, but they do not react to parasympathetic stimulation. Furthermore, the apocrine glands, despite their close embryological relationship to sweat glands, are controlled by adrenergic fibers rather than by cholinergic fibers and are controlled by the sympathetic centers of the central nervous system rather than by the parasympathetic centers.

3 **The Gastrointestinal System.** The gastrointestinal system has its own intrinsic set of nerves known as the *intramural plexus*. However, both parasympathetic and sympathetic stimulation can affect gastrointestinal activity—parasympathetic especially. Parasympathetic stimulation, in general, increases the overall degree of activity of the gastrointestinal tract by promoting peristalsis and relaxing the sphincters, thus allowing rapid propulsion of contents along the tract. This propulsive effect is associated with simultaneous increases in rates of secretion by many of the gastrointestinal glands, which was described earlier.

Normal function of the gastrointestinal tract is not very dependent on sympathetic stimulation. However, strong sympathetic stimulation inhibits peristalsis and increases the tone of the sphincters. The net result is greatly slowed propulsion of food through the tract.

4 **The Heart.** In general, sympathetic stimulation increases the overall activity of the heart. This is accomplished by increasing both the rate and force of the heartbeat. Parasympathetic stimulation causes mainly the opposite effects, decreasing the overall activity of

the heart. To express these effects in another way, sympathetic stimulation increases the effectiveness of the heart as a pump, whereas parasympathetic stimulation decreases its pumping capability.

Systemic Blood Vessels. Most blood vessels, especially those of the abdominal viscera and the skin of the limbs, are constricted by sympathetic stimulation. Parasympathetic stimulation generally has almost no effects on blood vessels but does dilate vessels in certain restricted areas such as in the blush area of the face. Under some conditions, the beta stimulatory function of the sympathetics causes vascular dilatation, especially when drugs have paralyzed the sympathetic alpha effects.

Effect of Sympathetic and Parasympathetic Stimulation on Arterial Pressure. The arterial pressure is caused by propulsion of blood by the heart and by resistance to flow of this blood through the blood vessels. In general, sympathetic stimulation increases both propulsion by the heart and resistance to flow, which can cause the pressure to increase greatly.

On the other hand, parasympathetic stimulation decreases the pumping by the heart, which lowers the pressure a moderate amount, though usually not nearly so much as the sympathetics can increase the pressure.

Effects of Sympathetic and Parasympathetic Stimulation on Other Functions of the Body. Because of the great importance of the sympathetic and parasympathetic control systems, these are discussed many times in this text in relation to a myriad of body functions that are not considered in detail here. In general, most of the entodermal structures, such as the ducts of the liver, the gallbladder, the ureter, the bladder, and the bronchi, are inhibited by sympathetic stimulation but excited by parasympathetic stimulation. Sympathetic stimulation also has metabolic effects, causing release of glucose from the liver, increase in blood glucose concentration, increase in glycogenolysis in muscle, increase in muscle strength, increase in basal metabolic rate, and increase in mental activity. Finally, the sympathetics and parasympathetics are involved in the execution of the male and female sexual acts, as will be explained in Chapters 80 and 81.

FUNCTION OF THE ADRENAL MEDULLAE

Stimulation of the sympathetic nerves to the adrenal medullae causes large quantities of epinephrine and norepinephrine to be released into the circulating blood, and these two hormones in turn are carried in the blood to all tissues of the body. On the average, approximately 80 per cent of the secretion is epinephrine and 20 per cent is norepinephrine, though the relative proportions of these change considerably under different physiological conditions.

The circulating epinephrine and norepinephrine have almost the same effects on the different organs as those caused by direct sympathetic stimulation, except that *the effects last 5 to 10 times as long* because these hormones are removed from the blood slowly.

The circulating norepinephrine causes constriction of essentially all the blood vessels of the body; it causes increased activity of the heart, inhibition of the gastrointestinal tract, dilation of the pupils of the eyes, and so forth.

Epinephrine causes almost the same effects as those caused by norepinephrine, but the effects differ in the following respects. First, epinephrine, because of its greater effect in stimulating the beta receptors, has a greater effect on cardiac activity than norepinephrine. Second, epinephrine causes only weak constriction of the blood vessels of the muscles in comparison with a much stronger constriction that results from norepinephrine. Since the muscle vessels represent a major segment of all vessels of the body, this difference is of special importance because norepinephrine greatly increases the total peripheral resistance and thereby greatly elevates arterial pressure, whereas epinephrine raises the arterial pressure to a lesser extent but increases the cardiac output considerably more because of its effect on the heart and veins.

A third difference between the action of epinephrine and norepinephrine relates to their effects on tissue metabolism. Epinephrine has several times as great a metabolic effect as norepinephrine. Indeed, the epinephrine secreted by the adrenal medullae can increase the metabolic rate of the body often to as much as 100 per cent above normal, in this way increasing the activity and excitability of the whole body. It also increases the rate of other metabolic activities, such as glycogenolysis in the liver and muscle and glucose release into the blood.

In summary, stimulation of the adrenal medullae causes the release of hormones that have almost the same effects throughout the body as direct sympathetic stimulation, except that the effects are greatly prolonged, up to a minute or two after the stimulation is over. The only significant differences are caused by the beta effects of the epinephrine in the secretion, which increase the rate of metabolism and cardiac output to a greater extent than is caused by direct sympathetic stimulation.

Value of the Adrenal Medullae to the Function of the Sympathetic Nervous System. Epinephrine and norepinephrine are almost always released by the adrenal medullae at the same time that the different organs are stimulated directly by the sympathetic nerves. Therefore, the organs are actually stimulated in two different ways simultaneously, directly by the sympathetic nerves and indirectly by the medullary hormones. The two means of stimulation support each other, and either can usually substitute for the other. For instance, destruction of the direct sympathetic pathways to the organs does not abrogate excitation of the organs because norepinephrine and epinephrine are still released into the circulating fluids and indirectly cause stimulation. Likewise, total loss of the two adrenal medullae usually has little significant effect on the operation of the sympathetic nervous system because the direct pathways can still perform almost all the neces

ary duties. Thus, the dual mechanism of sympa-
hetic stimulation provides a safety factor, one
mechanism substituting for the other when the
econd is missing.

Another important value of the adrenal medul-
ae is the capability of epinephrine and norepi-
hrine to stimulate structures of the body that
not innervated by direct sympathetic fibers.
For instance, the metabolic rate of every cell of
body is increased by these hormones, especially
epinephrine, even though only a small propor-
tion of all the cells in the body are innervated
rectly by sympathetic fibers.

STIMULUS RATE TO DEGREE OF SYMPATHETIC AND PARASYMPATHETIC EFFECT

A special difference between the autonomic ner-
system and the skeletal nervous system is
the *low frequency of stimulation required for full*
vation of autonomic effectors. In general, only
one impulse every second or so suffices to maintain
sympathetic or parasympathetic effect, and
activation occurs when the nerve fibers dis-
charge 10 to 20 times per second. This compares
with full activation in the skeletal nervous system
about 50 to 200 impulses per second.

SYMPATHETIC AND PARASYMPATHETIC

The sympathetic and parasympathetic systems are continually active, and the basal rates of activity are known, respectively, as *sympathetic tone* or *parasympathetic tone*.

The value of tone is that it allows a single nervous system to *increase or to decrease the activity* of a stimulated organ. For instance, sympathetic tone normally keeps almost all the blood vessels of the body constricted to approximately half their maximum diameter. By increasing the degree of sympathetic stimulation, the vessels can be constricted even more, but, on the other hand, by inhibiting the normal tone, the vessels can be dilated. If it were not for the continual sympathetic tone, the sympathetic system could cause only vasoconstriction, never vasodilatation.

Another interesting example of tone is that of the parasympathetics in the gastrointestinal tract. Surgical removal of the parasympathetic supply to the gut by cutting the vagi can cause serious and prolonged gastric and intestinal "atony," thus illustrating that in normal function the parasympathetic tone to the gut is strong. This tone can be decreased by the brain, thereby inhibiting gastrointestinal motility, or it can be increased, thereby promoting increased gastrointestinal activity.

Tone Caused by Basal Secretion of Epinephrine and Norepinephrine by the Adrenal Medullae. The normal resting rate of secretion by the adrenal medullae is about 0.2 μgm/kg/min of epinephrine and about 0.05 μgm/kg/min of norepinephrine. These quantities are considerable—indeed, enough to maintain the blood pressure almost up to the normal value even if all direct sympathetic pathways to the cardiovascular system are removed. Therefore, it is obvious that much of the overall tone of the sympathetic nervous system results from basal secretion of epinephrine and norepinephrine in addition to the tone resulting from direct sympathetic stimulation.

Effect of Loss of Sympathetic or Parasympathetic Tone Following Denervation. Immediately after a sympathetic or parasympathetic nerve is cut, the innervated organ loses its sympathetic or parasympathetic tone. In the case of the blood vessels, for instance, cutting the sympathetic nerves results immediately in almost maximal vasodilatation. However, over minutes, hours, days, or weeks, the *intrinsic tone* in the smooth muscle of the vessels increases, usually restoring almost normal vasoconstriction.

Essentially the same events occur in most effector organs whenever sympathetic or parasympathetic tone is lost. That is, intrinsic compensation soon develops to return the function of the organ almost to its normal basal level. However, in the parasympathetic system, the compensation sometimes requires many months. For instance, loss of parasympathetic tone to the heart *increases the heart rate to 180 beats per minute* in a dog, and this will still be about 120 beats six months later.

DENERVATION SUPERSENSITIVITY OF SYMPATHETIC AND PARASYMPATHETIC ORGANS FOLLOWING DENERVATION

During the first week or so after a sympathetic or parasympathetic nerve is destroyed, the innervated organ becomes more and more sensitive to injected norepinephrine or acetylcholine, respectively. This effect is illustrated in Figure 57–4, the blood flow in the forearm before removal of the sympathetics is about 200 ml per minute, and a test dose of norepinephrine causes only a slight depression in flow. Then the stellate ganglion is removed, and normal sympathetic tone is lost. At first, the blood flow rises markedly because of the lost vascular tone, but over a period of days to weeks the blood flow returns almost to normal because of progressive increase in intrinsic tone of the vascular musculature itself, thus compensating for the loss of sympathetic tone. Another test dose of norepinephrine is then administered and the blood flow decreases much more than before, illustrating that the blood vessels become about two to four

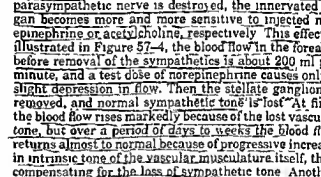


Figure 57–4. Effect of sympathectomy on blood flow in the arm, and the effect of a test dose of norepinephrine before and after sympathectomy, showing sensitization of the vasculature to norepinephrine.

times as responsive to norepinephrine as previously. This phenomenon is called *denervation supersensitivity*. It occurs in both sympathetic and parasympathetic organs and to a far greater extent in some organs than in others, often increasing the response as much as tenfold or more.

Mechanism of Denervation Supersensitivity. The cause of denervation supersensitivity is only partially known. Part of the answer is that the number of receptors in the postsynaptic membranes of the effector cells increases—sometimes manyfold—when norepinephrine or acetylcholine are no longer released at the synapses, a process called "up-regulation" of the receptors. Therefore, when these hormones appear in the circulating blood, the effector reaction is vastly enhanced.

THE AUTONOMIC REFLEXES

Many of the visceral functions of the body are regulated by *autonomic reflexes*. Throughout this text the functions of these reflexes are discussed in relation to individual organ systems, but, to illustrate their importance, a few are presented here briefly.

Cardiovascular Autonomic Reflexes. Several reflexes in the cardiovascular system help to control the arterial blood pressure, cardiac output, and heart rate. One of these is the *baroreceptor reflex*, which was described in Chapter 22 along with other cardiovascular reflexes. Briefly, stretch receptors called *baroreceptors* are located in the walls of the major arteries, including the carotid arteries and the aorta. When these become stretched by high pressure, signals are transmitted to the brain stem, where they inhibit the sympathetic impulses to the heart and blood vessels, which allows the arterial pressure to fall back toward normal.

The Gastrointestinal Autonomic Reflexes. The uppermost part of the gastrointestinal tract and also the rectum are controlled principally by autonomic reflexes. For instance, the smell of appetizing food initiates signals from the nose to the vagal, glossopharyngeal, and salivary nuclei of the brain stem. These in turn transmit signals through the parasympathetic nerves to the secretory glands of the mouth and stomach, causing secretion of digestive juices even before food enters the mouth. And when fecal matter fills the rectum at the other end of the alimentary canal, sensory impulses initiated by stretching the rectum are sent to the sacral portion of the spinal cord, and a reflex signal is retransmitted through the parasympathetics to the distal parts of the colon; these result in strong peristaltic contractions that empty the bowel.

Other Autonomic Reflexes. Emptying of the bladder is controlled in the same way as emptying the rectum; stretching of the bladder sends impulses to the sacral cord, and this in turn causes contraction of the bladder as well as relaxation of the urinary sphincters, thereby promoting micturition. Important are the sexual reflexes which are initiated both by psychic stimuli from the brain and stimuli from the sexual organs. Impulses from these sources converge on the sacral cord and, in the male, result, first, in erection, mainly a parasympathetic function, and then in ejaculation, a sympathetic function.

Other autonomic reflexes include reflex contributions to the regulation of pancreatic secretion, gallbladder emptying, renal excretion of urine, sweating, blood glucose concentration, and many other visceral functions, all of which are discussed in detail at other points in this text.

STIMULATION OF DISCRETE ORGANS IN SOME INSTANCES, MASS STIMULATION IN OTHER INSTANCES, BY THE SYMPATHETIC AND PARASYMPATHETIC SYSTEMS

The Sympathetic System. In many instances the sympathetic nervous system discharges almost as a complete unit, a phenomenon called *mass discharge*. This frequently occurs when the hypothalamus becomes activated by fright or fear or severe pain. The result is a widespread reaction throughout the body called the *alarm or stress response*, which we shall discuss shortly.

However, at other times sympathetic activation occurs in isolated portions of the system. The most important of these are: (1) In the process of heat regulation, the sympathetics control sweating and blood flow in the skin without affecting other organs innervated by the sympathetics. (2) During muscular activity in some animals, cholinergic vasodilator fibers of the skeletal muscles are stimulated independently of all the remainder of the sympathetic system. (3) Many "local reflexes" involving the spinal cord but usually not the higher nervous centers affect local areas. For instance, heating a local skin area causes local vasodilation and enhanced local sweating, while cooling causes the opposite effects. (4) Many of the sympathetic reflexes that control gastrointestinal functions are very discrete, operating sometimes by way of nerve pathways that do not even enter the spinal cord, merely passing from the gut to the sympathetic ganglia and then back to the gut through the sympathetic nerves to control motility or secretory activity.

The Parasympathetic System. In contrast to the sympathetic system, most control functions of the parasympathetic system are very specific. For instance, parasympathetic cardiovascular reflexes usually act only on the heart to increase or decrease its rate of beating. Likewise, parasympathetic reflexes frequently cause secretion mainly in the mouth or, in other instances, secretion mainly in the stomach glands. Finally, the rectal emptying reflex does not affect other parts of the bowel to a major extent.

Yet there is often association between close

ied parasympathetic functions. For instance, ough salivary secretion can occur independently gastric secretion, these two often also occur gether, and pancreatic secretion frequently oc- at the same time. Also, the rectal emptying flex often initiates a bladder emptying reflex, sulting in simultaneous emptying of both the adder and rectum. Conversely, the bladder emp- ing reflex can help initiate rectal emptying.

‹LARM" OR "STRESS" RESPONSE OF THE 'MPATHETIC NERVOUS SYSTEM

When large portions of the sympathetic nervous stem discharge at the same time—that is, a mass scharge—this increases in many different ways e capability of the body to perform vigorous uscle activity. Let us quickly summarize these ays:

1. increased arterial pressure
2. increased blood flow to active muscles concur- nt with decreased blood flow to organs that are t needed for rapid activity
3. increased rates of cellular metabolism roughout the body
4. increased blood glucose concentration
5. increased glycolysis in muscle
6. increased muscle strength
7. increased mental activity
8. increased rate of blood coagulation

The sum of these effects permits the person to erform far more strenuous physical activity than ould otherwise be possible. Since it is physical *ress* that usually excites the sympathetic system, is frequently said that the purpose of the sym- athetic system is to provide extra activation of e body in states of stress: this is often called the mpathetic *stress response*.

The sympathetic system is also strongly acti- ated in many emotional states. For instance, in e state of *rage*, which is elicited mainly by imulating the hypothalamus, signals are trans- itted downward through the reticular formation nd spinal cord to cause massive sympathetic dis- harge, and all of the sympathetic events listed bove ensue immediately. This is called the sym- athetic *alarm reaction*. It is also frequently called e *fight or flight reaction* because an animal in is state decides almost instantly whether to tand and fight or to run. In either event, the ympathetic alarm reaction makes the animal's ubsequent activities extremely vigorous.

IEDULLARY, PONTINE, AND IESENCEPHALIC CONTROL OF THE UTONOMIC NERVOUS SYSTEM

Many areas in the reticular substance of the iedulla, pons, and mesencephalon, as well as iany special nuclei (see Fig. 57–5), control differ- nt autonomic functions such as arterial pressure, ieart rate, glandular secretion in the upper part f the gastrointestinal tract, gastrointestinal per-

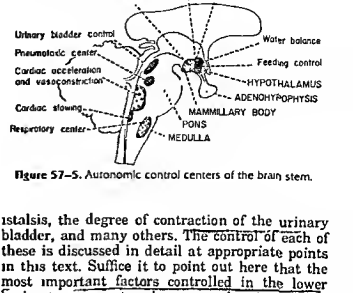


Figure 57–5. Autonomic control centers of the brain stem.

istalsis, the degree of contraction of the urinary bladder, and many others. The control of each of these is discussed in detail at appropriate points in this text. Suffice it to point out here that the most important factors controlled in the lower brain stem are arterial pressure, heart rate, and respiration. Indeed, transection of the brain stem at the midpontine level allows normal basal control of arterial pressure to continue as before but prevents its modulation by higher nervous centers, particularly the hypothalamus. On the other hand, transection immediately below the medulla causes the arterial pressure to fall to about one-half normal for several hours or several days after the transection.

Closely associated with the cardiovascular regulatory centers in the medulla are the medullary and pontine centers for regulation of respiration discussed in detail in Chapter 42. Though this is not considered to be an autonomic function, it is one of the *involuntary* functions of the body.

Control of Lower Brain Stem Autonomic Centers by Higher Areas. Signals from the hypothalamus and even from the cerebrum can affect the activities of almost all the lower brain stem autonomic control centers. For instance, stimulation in appropriate areas of the hypothalamus can activate the medullary cardiovascular control centers strongly enough to increase the arterial pressure to more than double normal. Likewise, other hypothalamic centers can control body temperature, increase or decrease salivation and gastrointestinal activity, or cause bladder emptying. To some extent, therefore, the autonomic centers in the lower brain stem act as relay stations for control activities initiated at higher levels of the brain.

In the previous chapter it was also pointed out that many of the behavioral responses of an animal are mediated through the hypothalamus, reticular formation, and the autonomic nervous system. Indeed, the higher areas of the brain can alter the function of the whole autonomic nervous system or of portions of it strongly enough to cause severe autonomic-induced disease, such as peptic ulcer, constipation, heart palpitation, and even heart attacks.

X

THE SPECIAL SENSES

58 ■ The Eye: I. Optics of Vision

59 ■ The Eye: II. Receptor and Neural Function of the Retina

60 ■ The Eye: III. Central Neurophysiology of Vision

61 ■ The Sense of Hearing

62 ■ The Chemical Senses—Taste and Smell

58

The Eye: I. Optics of Vision

PHYSICAL PRINCIPLES OF OPTICS

Before it is possible to understand the optical system of the eye, the student must be thoroughly familiar with the basic physical principles of optics, including the physics of refraction, a knowledge of focusing, depth of focus, and so forth. Therefore, in the present study of the optics of the eye, a brief review of these physical principles is first presented, and then the optics of the eye is discussed.

REFRACTION OF LIGHT

3×10^8 m/s

The Refractive Index of a Transparent Substance. Light rays travel through air at a velocity of approximately 300,000 kilometers per second but much slower through transparent solids and liquids. The refractive index of a transparent substance is the *ratio* of the velocity of light in air to that in the substance. Obviously, the refractive index of air itself is 1.00.

If light travels through a particular type of glass at a velocity of 200,000 kilometers per second, the refractive index of this glass is 300,000 divided by 200,000, or 1.50.

Refraction of Light Rays at an Interface Between Two Media with Different Refractive Indices. When light waves traveling forward in a beam, as shown in the upper part of Figure 58–1, strike an interface that is perpendicular to the beam, the waves enter the second refractive medium without deviating in their course. The only effect that occurs is decreased velocity of transmission and shorter wavelength. On the other hand, as illustrated in the lower part of Figure 58–1, if the light waves strike an angulated interface, the light waves bend if the refractive indices of the two media are different from each other. In this particular figure the light waves are leaving air, which has a refractive index of 1.00, and are entering a block of glass having a refractive index of 1.50. When the beam first strikes the angulated interface, the lower edge of the beam enters the glass ahead of the upper edge. The wave front in the upper portion of the beam continues to travel at a velocity of 300,000 kilometers per second while that which has entered the glass travels at a velocity of 200,000 kilometers per second. This causes the upper portion of the wave front to move ahead of the lower portion so that the wave front is no longer vertical but is angulated to the right. Because *the direction in which light travels is always perpendicular to the plane of wave front*, the direction of travel of the light beam bends downward.

The bending of light rays at an angulated interfa known as *refraction*. Note particularly that the de of refraction increases as a function of (1) the rat the two refractive indices of the two transparent m and (2) the degree of angulation between the inter and the entering wave front.

APPLICATION OF REFRACTIVE PRINCIPLES TO LENSES

The Convex Lens—Focusing of Light Rays. Fi 58–2 shows parallel light rays entering a convex l The light rays passing through the center of the l strike the lens exactly perpendicular to the lens sur and therefore pass through the lens without being fracted at all. Toward either edge of the lens, howe the light rays strike a progressively more angula interface. Therefore, the outer rays bend more and m toward the center. Half the bending occurs when rays enter the lens and half as they exit from opposite side. (At this time the student should pa and analyze why the rays still bend toward the cen upon leaving the lens.)

Finally, if the lens is ground with exactly the pro curvature, parallel light rays passing through each p

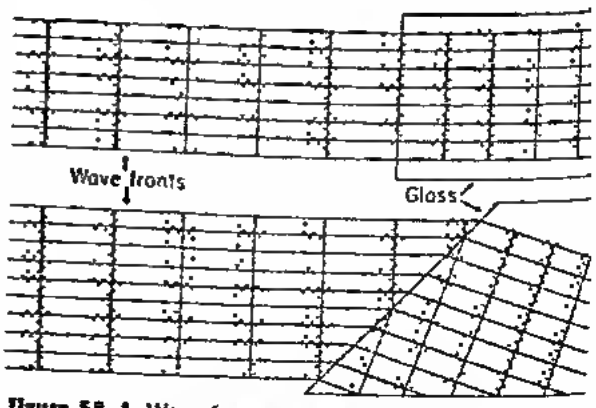


Figure 58–1. Wave fronts entering *(top)* a glass surface pe pendicular to the light rays and *(bottom)* a glass surface angulate to the light rays. This figure illustrates that the distance betwee waves after they enter the glass is shortened to approximatel two thirds that in air. It also illustrates that light rays striking a angulated glass surface are refracted.

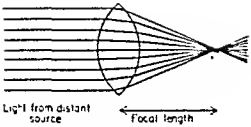


Figure 58–2. Bending of light rays at each surface of a convex spherical lens, showing that parallel light rays are focused to a point focus.

of the lens will be bent exactly enough so that all the rays will pass through a single point, which is called the *focal point*.

The Concave Lens. Figure 58–3 shows the effect of a concave lens on parallel light rays. The rays that enter the very center of the lens strike an interface that is absolutely perpendicular to the beam and, therefore, do not refract at all. The rays at the edge of the lens enter the lens ahead of the rays toward the center. This is opposite to the effect in the convex lens, and it causes the peripheral light rays to *diverge* away from the light rays that pass through the center of the lens.

Thus, the concave lens *diverges* light rays, whereas the convex lens *converges* light rays.

Spherical Versus Cylindrical Lenses. Figure 58–4 illustrates both a *convex spherical* lens and a *convex cylindrical* lens. Note that the cylindrical lens bends light rays from the two sides of the lens but not from either the top or the bottom. Therefore, parallel light rays are bent to a focal *line*. On the other hand, light rays that pass through the spherical lens are refracted at all edges of the lens toward the central ray, and all the rays come to a *focal point*.

The cylindrical lens is well illustrated by a test tube full of water. If the test tube is placed in a beam of sunlight and a piece of paper is brought progressively closer to the tube, a certain distance will be found at which the light rays come to a *focal line*. On the other hand, the spherical lens is illustrated by an ordinary magnifying glass. If such a lens is placed in a beam of sunlight and a piece of paper is brought progressively closer to the lens, the light rays will impinge on a common focal point at an appropriate distance.

Concave cylindrical lenses *diverge* light rays in only one plane in the same manner that *convex* cylindrical lenses *converge* light rays in one plane.

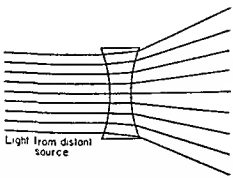


Figure 58–3. Bending of light rays at each surface of a concave spherical lens, illustrating that parallel light rays are diverged by a concave lens.

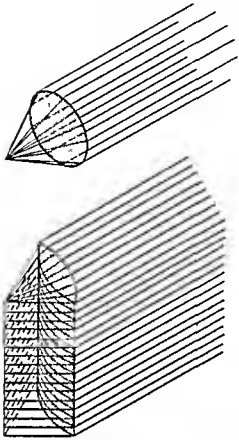


Figure 58–4. *Top:* Point focus of parallel light rays by a spherical convex lens. *Bottom:* Line focus of parallel light rays by a cylindrical convex lens.

Figure 58–5 shows two convex cylindrical lenses at right angles to each other. The vertical cylindrical lens causes convergence of the light rays that pass through the two sides of the lens. The horizontal lens converges the top and bottom rays. Thus, all the light rays come

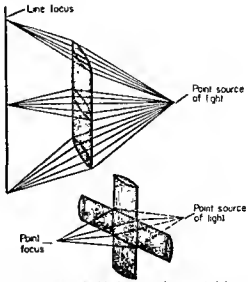


Figure 58–5. Two cylindrical convex lenses at right angles to each other, illustrating that one lens converges light rays in one plane and the other lens converges light rays in the plane at right angles. The two lenses combined give the same point focus as that obtained with a spherical convex lens.

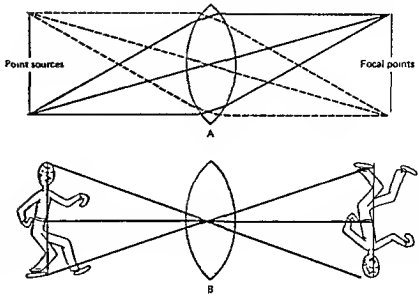


Figure 58–7. A. Two point sources of light focused two separate points on the opposite side of the ns. B. Formation of an image by a convex spherical ns.

Note particularly that concave lenses can "neutrala" the refractive power of convex lenses. Thus, placing 1 diopter concave lens immediately in front of a 1 iopter convex lens results in a lens system with zero efractive power.

The strengths of cylindrical lenses are computed in ie same manner as the strengths of spherical lenses. If cylindrical lens focuses parallel light rays to a line ocus 1 meter beyond the lens, it has a strength of +1 iopter. On the other hand, if a cylindrical lens of a oncave type *diverges* light rays as much as a +1 diopter ylindrical lens *converges* them, it has a strength of −1 iopter.

HE OPTICS OF THE EYE

HE EYE AS A CAMERA

The eye, as illustrated in Figure 58–9, is optically equivalent to the usual photographic camera, or it has a lens system, a variable aperture system the pupil), and a retina that corresponds to the ilm. The lens system of the eye is composed of our refractive interfaces: (1) the interface between iir and the anterior surface of the cornea, (2) the interface between the posterior surface of the cornea and the aqueous humor, (3) the interface between the aqueous humor and the anterior surface of the crystalline lens of the eye, and (4) the interface between the posterior surface of the lens and the vitreous humor. *The refractive index of* air is 1; the cornea, 1.38; the aqueous humor, 1.33; the crystalline lens (on the average), 1.40; and the vitreous humor, 1.34.

The Reduced Eye. If all the refractive surfaces of the eye are algebraically added together and then considered to be one single lens, the optics of the normal eye may be simplified and represented schematically as a "reduced eye." This is useful in simple calculations. In the reduced eye, a single lens is considered to exist with its central point 17 mm in front of the retina and to have a total refractive power of approximately 59 diopters when the lens is accommodated for distant vision.

Most of the refractive power of the eye is provided not by the crystalline lens but instead by the anterior surface of the cornea. The principal reason for this is that the refractive index of the cornea is markedly different from that of air.

On the other hand, the total refractive power of the crystalline lens of the eye, as it normally lies in the eye surrounded by fluid on each side, is only 15 diopters, about one fourth the total refractive

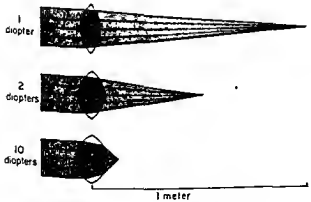


Figure 58–8. Effect of lens strength on the focal distance

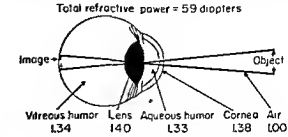


Figure 58–9. The eye as a camera. The numbers are the refractive indices.

power of the eye's lens system. If this lens were removed from the eye and then surrounded by air, its refractive power would be about six times as great. The reason for this difference is that the fluids surrounding the lens have refractive indices not greatly different from the refractive index of the lens itself, and the smallness of the differences greatly decreases the amount of light refraction at the lens interfaces. But the importance of the crystalline lens is that its curvature can be increased markedly to provide "accommodation," which will be discussed later in the chapter.

Formation of an Image on the Retina. In exactly the same manner that a glass lens can focus an image on a sheet of paper, the lens system of the eye can focus an image on the retina. The image is inverted and reversed with respect to the object. However, the mind perceives objects in the upright position despite the upside-down orientation on the retina because the brain is trained to consider an inverted image as the normal.

THE MECHANISM OF ACCOMMODATION

The refractive power of the crystalline lens of the eye can be voluntarily increased from 15 diopters to approximately 29 diopters in young children; this is a total "accommodation" of 14 diopters. To do this, the shape of the lens is changed from that of a moderately convex lens to that of a very convex lens. The mechanism of this is the following:

In the young person, the lens is composed of a strong elastic capsule filled with viscous, proteinaceous, but transparent fibers. When the lens is in a relaxed state, with no tension on its capsule, it assumes a spherical shape, owing entirely to the elasticity of the lens capsule. However, as illustrated in Figure 58–10, approximately 70 ligaments attach radially around the lens, pulling the lens edges toward the edge of the choroid. These ligaments are constantly tensed by the elastic pull of their attachments to the choroid, and the tension on the ligaments causes the lens to remain relatively flat under normal resting conditions of the eye. At the insertions of the ligaments in the choroid is the ciliary muscle, which has two sets of smooth muscle fibers, the *meridional fibers* and the *circular fibers*. The meridional fibers extend from the corneoscleral junction to the insertions of the lens ligaments in the choroid approximately 2 to 3 mm behind the corneoscleral junction. When these muscle fibers contract, the insertions of the ligaments are pulled forward, thereby releasing a certain amount of tension on the crystalline lens. The circular fibers are arranged circularly all the way around the eye so that when they contract a sphincter-like action occurs, decreasing the diameter of the circle of ligament attachments and allowing the ligaments to pull less on the lens capsule.

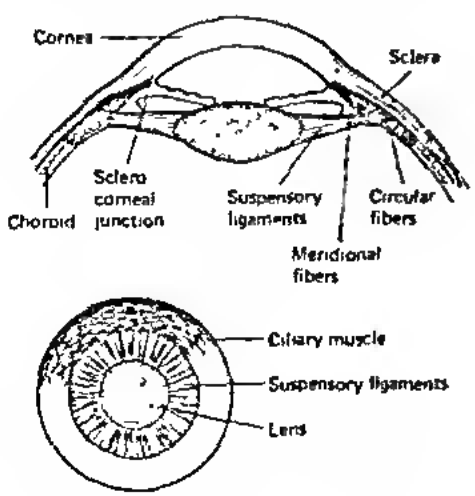


Figure 58–10. Mechanism of accommodation (focusing)

Thus, contraction of both sets of smooth muscle fibers in the ciliary muscle relaxes the ligaments to the lens capsule, and the lens assumes a more spherical shape, like that of a balloon, because of the natural elasticity of its capsule. When the ciliary muscle is completely relaxed, the dioptric strength of the lens is as weak as it can become. On the other hand, when the ciliary muscle contracts as strongly as possible, the dioptric strength of the lens becomes maximal.

Autonomic Control of Accommodation. The ciliary muscle is controlled almost entirely by the parasympathetic nervous system. Stimulation of the parasympathetic nerves to the eye contracts the ciliary muscle, which in turn relaxes the lens ligaments and increases the refractive power. With an increased refractive power, the eye is capable of focusing on objects nearer at hand than when the eye has less refractive power. Consequently, as a distant object moves toward the eye, the number of parasympathetic impulses impinging on the ciliary muscle must be progressively increased for the eye to keep the object constantly in focus. (Sympathetic stimulation has a weak effect in relaxing the ciliary muscle but this plays almost no role in the normal accommodation mechanism, the neurology of which will be discussed in Chapter 60.)

Presbyopia. As a person grows older, the lens loses its elastic nature and becomes a relatively solid mass, partly because of progressive denaturation of the lens proteins. Therefore, the ability of the lens to assume a spherical shape progressively decreases, and the power of accommodation decreases from approximately 14 diopters in the young child to less than 2 diopters at the age of 45 to 50. Thereafter, the lens may be considered to be almost totally nonaccommodating, a condition known as "presbyopia."

Once a person has reached the state of presbyopia, each eye remains focused permanently at an almost constant distance, this distance depends on

the physical characteristics of each individual's eyes. Obviously, the eyes can no longer accommodate for both near and far vision. Therefore, to see clearly both in the distance and nearby, an older person must wear bifocal glasses with the upper segment normally focused for far-seeing and the lower segment focused for near-seeing.

THE PUPILLARY APERTURE

A major function of the iris is to increase the amount of light that enters the eye during darkness and to decrease the light in bright light. The reflexes for controlling this mechanism will be considered in the discussion of the neurology of the eye in Chapter 60. The amount of light that enters the eye through the pupil is proportional to the *area* of the pupil or to the *square of the diameter* of the pupil. The pupil of the human eye can become as small as approximately 1.5 mm and as large as 8 mm in diameter. Therefore, the quantity of light entering the eye may vary approximately 30 times as a result of changes in pupillary aperture.

Depth of Focus of the Lens System of the Eye. Figure 58–11 illustrates two separate eyes that are exactly alike except for the diameters of the pupillary apertures. In the upper eye the pupillary aperture is small, and in the lower eye the apertura is large. In front of each of these two eyes are two small point sources of light, and light from each passes through the pupillary aperture and focuses on the retina. Consequently, in both eyes the retina sees two spots of light in perfect focus. It is evident from the diagrams, however, that if the retina is moved forward or backward to an out-of-focus position, the size of each spot will not change much in the upper eye, but in the lower eye the size of each spot will increase greatly, becoming a "blur circle." In other words, the upper lens system has far greater *depth of focus* than the bottom lens system. When a lens system has great depth of focus, the retina can be considerably displaced from the focal plane and still the image remains in sharp focus; whereas, when a lens system has a shallow depth of focus, moving the retina only slightly away from the focal plane causes extreme blurring.

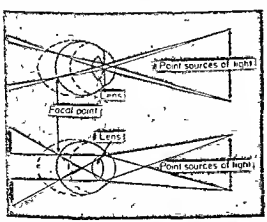


Figure 58–11. Effect of small and large pupillary apertures on the depth of focus.

The greatest possible depth of focus occurs when the pupil is extremely small. The reason for this is that with a very small aperture all light rays must pass through the center of the lens, and the centralmost rays are always in focus, as was explained earlier.

ERRORS OF REFRACTION

Emmetropia. As shown in *Figure 58–12, the eye is* considered to be normal, or "emmetropic," if parallel light rays from distant objects are in sharp focus on the retina when the ciliary muscle is completely relaxed. This means that the emmetropic eye can see all distant objects clearly, with its ciliary muscle completely relaxed, but to focus objects at close range it must contract *its ciliary muscle and thereby provide various degrees* of accommodation.

Hypermetropia (Hypeopia). Hypermetropia, which is also known as "far-sightedness," is usually due either to an eyeball that is too short or occasionally to a lens system that is too weak when the ciliary muscle is relaxed. In this condition, as seen in the middle panel *of Figure 58–12, parallel light rays are not bent suffi*ciently by the lens system to come to a focus by the time they reach the retina. To overcome this abnormality, the ciliary muscle may contract to increase the strength of the lens. Therefore, the far-sighted person is capable, by using his mechanism of accommodation, of focusing distant objects on the retina. If he has used only a small amount of strength in his ciliary muscle to accommodate for the distant objects, then he still has much accommodative power left, and objects closer and closer to the eye can also be focused sharply until the ciliary muscle has contracted to its limit. The distance of the object

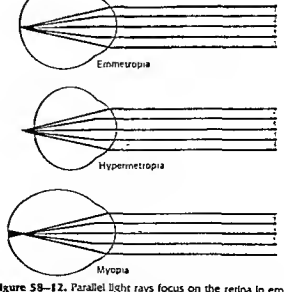


Figure 58–12. Parallel light rays focus on the retina in emmetropia, behind the retina in hypermetropia, and in front of the retina in myopia.

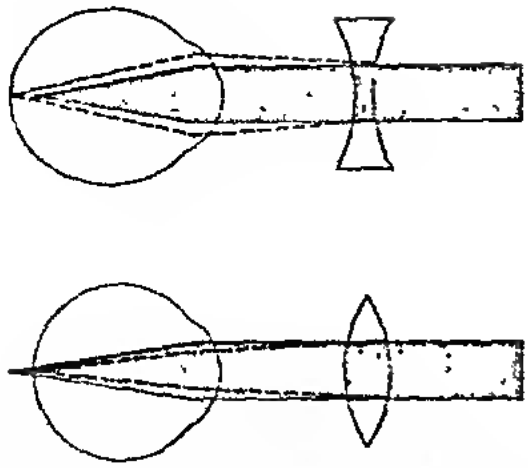


Figure 58–13. Correction of myopia with a concave lens, and correction of hypermetropia with a convex lens.

away from the eye at this point is known as the "near point" of vision.

In old age, when the lens becomes presbyopic, the far-sighted person often is not able to accommodate his or her lens sufficiently to focus even distant objects, much less to focus near objects

Myopia. In myopia, or "near-sightedness," when the ciliary muscle is completely relaxed, the light rays coming from distant objects are focused in front of the retina, as shown in the lower panel of Figure 58–12. This is usually due to too long an eyeball but it can occasionally result from too much refractive power of the lens system of the eye

No mechanism exists by which the eye can decrease the strength of its lens to less than that which exists when the ciliary muscle is completely relaxed Therefore, the myopic person has no mechanism by which he can ever focus distant objects sharply on his retina However, as an object comes nearer to his eye it finally comes near enough that its image will focus on the retina Then, when the object comes still closer to the eye, the person can use his mechanism of accommodation to keep the image focused clearly Therefore, a myopic person has a definite limiting "far point" for clear vision as well as a "near point"

Correction of Myopia and Hypermetropia by Use of Lenses. It will be recalled that light rays passing through a concave lens diverge. Therefore, if the refractive surfaces of the eye have too much refractive power, as in myopia, some of this excessive refractive power can be neutralized by placing in front of the eye a concave spherical lens, which will diverge rays On the other hand, in a person who has hypermetropia—that is, someone who has too weak a lens system—the abnormal vision can be corrected by adding refractive power with a convex lens in front of the eye These corrections are illustrated in Figure 58–13. One usually determines the strength of the concave or convex lens needed for clear vision by "trial and error"—that is, by trying first a strong lens and then a stronger or weaker lens until the one that gives the best visual acuity is found

Astigmatism

Astigmatism is a refractive error of the lens system, caused usually by an oblong shape of the cornea or, 'y, an oblong shape of the lens. A lens surface like the side of an egg lying edgewise to the incoming light would be an example of an astigmatic lens. The degree of curvature in the plane through the long axis of the egg is not nearly so great as the degree of curvature in the plane through the short axis. Because the curvature of the astigmatic lens along one plane is less than the curvature along the other plane, light rays striking the peripheral portions of the lens in one plane are not bent nearly so much as are rays striking the peripheral portions of the other plane.

This is illustrated in Figure 58–14, which shows rays of light emanating from a point source and passing through an oblong, astigmatic lens. The light rays in the vertical plane, indicated by plane BD, are refracted greatly by the astigmatic lens because of the greater curvature in the vertical direction than in the horizontal direction However, the light rays in the horizontal plane, indicated by plane AC, are bent not nearly so much as the light rays in the vertical plane. It is obvious therefore, that the light rays passing through an astigmatic lens do not all come to a common focal point because the light rays passing through one plane of the lens focus far in front of those passing through the other plane.

The accommodative powers of the eyes can never compensate for astigmatism because, during accommodation, the curvature of the eye lens changes equally in both planes Therefore, when the accommodation corrects the refractive error in one plane, the error in the other plane is not corrected That is, each of the two planes requires a different degree of accommodation to be corrected, so that the two planes are never corrected at the same time Thus, vision never occurs with a sharp focus

Correction of Astigmatism with a Cylindrical Lens. One may consider an astigmatic eye as having a lens system made up of two cylindrical lenses of different strengths and placed at right angles to each other Therefore, to correct for astigmatism the usual procedure is to find a spherical lens by "trial and error" that corrects the focus in one of the two planes of the astigmatic lens Then an additional cylindrical lens is used to correct the error in the remaining plane. To do this, both the *axis* and the *strength* of the required cylindrical lens must be determined.

There are several methods for determining the axis of the abnormal cylindrical component of the lens system of an eye. One of these methods is based on the use

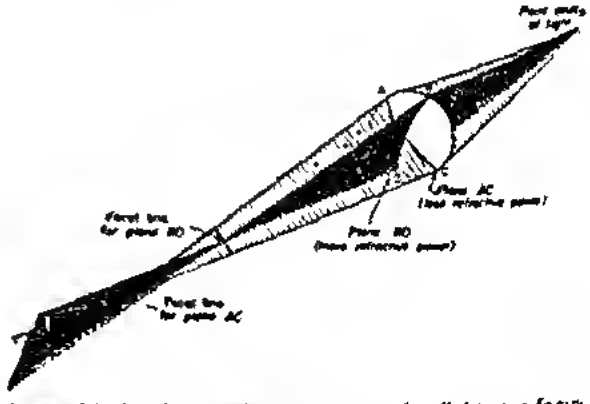


Figure 58–14. Astigmatism, illustrating that light rays focus one focal distance in one focal plane and at another focal distance in the plane at right angles.

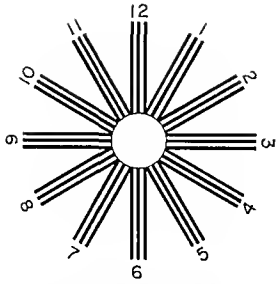


Figure 58–15. Chart composed of parallel black bars for determining the axis of astigmatism.

parallel black bars as shown in Figure 58–15. Some of these parallel bars are vertical, some horizontal, and some at various angles to the vertical and horizontal axes. After placing various spherical lenses in front of the astigmatic eye by trial and error, a strength of lens usually will be found that will cause sharp focus of one set of these parellel bars on the retina of the astigmatic eye but will cause fuzziness of the set of bars at right angles to the sharp bars.

It can be shown from the physical principles of optics discussed earlier in this chapter that the axis of the *out-of-focus* cylindrical component of the optical system is parallel to the bars that are fuzzy. Once this axis is found, the examiner tries progressively stronger and weaker positive or negative cylindrical lenses, the axes of which are placed parallel to the out-of-focus bars, until the patient sees all the crossed bars with equal clarity. When this has been accomplished, the examiner directs the optician to grind a special lens having the spherical correction as well as the cylindrical correction at the appropriate axis.

Correction of Optical Abnormalities by Use of Contact Lenses

In recent years, either glass or plastic contact lenses have been fitted snugly against the anterior surface of the cornea. These lenses are held in place by a thin layer of tears that fill the space between the contact lens and the anterior eye surface.

A special feature of the contact lens is that it nullifies almost entirely the refraction that normally occurs at the anterior surface of the cornea. The reason for this is that the tears between the contact lens and the cornea have a refractive index almost equal to that of the cornea so that no longer does the anterior surface of the cornea play a significant role in the eye's optical system. Instead, the anterior surface of the contact lens now plays the major role and its posterior surface a minor role. Thus, the refraction of this lens now substitutes for the cornea's usual refraction. This is especially important in persons whose eye refractive errors are caused by an abnormally shaped cornea, such as persons who have an odd-shaped, bulging cornea—a condition called *keratoconus*. Without the contact lens the bulging cornea causes such severe abnormality of vision that almost no glasses can correct the vision satisfactorily; when a contact lens is used, however, the corneal refraction is neutralized, and normal refraction by the anterior surface of the contact lens is substituted in its place.

The contact lens has several other advantages as well, including (1) the lens turns with the eye and gives a broader field of clear vision than do usual glasses, and (2) the contact lens has little effect on the size of the object that the person sees through the lens; on the other hand, lenses placed several centimeters in front of the eye do affect the size of the image in addition to correcting the focus.

Cataracts

Cataracts are an especially common eye abnormality that occurs in older people. A cataract is a cloudy or opaque area in the lens. In the early stage of cataract formation the proteins in the lens fibers immediately beneath the capsule become denatured. Later, these same proteins coagulate to form opaque areas in place of the normal transparent protein fibers of the lens. Finally, in still later stages, calcium is often deposited in the coagulated proteins, thus further increasing the opacity.

When a cataract has obscured light transmission so greatly that it seriously impairs vision, the condition can be corrected by surgical removal of the entire lens. When this is done, however, the eye loses a large portion of its refractive power, which must be replaced by a powerful convex lens (about +15 diopters) in front of the eye, or an artificial lens may be implanted inside the eye in place of the removed lens.

VISUAL ACUITY

Theoretically, light from a distant point source, when focused on the retina, should be infinitely small. However, since the lens system of the eye is not perfect, such a retinal spot ordinarily has a total diameter of about 11 microns even with maximal resolution of the optical system. However, it is brightest in its very center and shades off gradually toward the edges, as illustrated by the two point images in Figure 58–16.

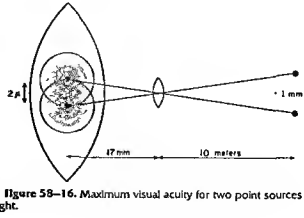


Figure 58–16. Maximum visual acuity for two point sources of light.

The average diameter of cones in the fovea of the retina, the central part of the retina where vision is most highly developed, is approximately 1.5 microns, which is one seventh the diameter of the spot of light. Nevertheless, since the spot of light has a bright center point and shaded edges, a person can distinguish two separate points if their centers lie approximately 2 microns apart on the retina, which is slightly greater than the width of a foveal cone. This discrimination between points is illustrated in Figure 58–16.

The maximum visual acuity of the human eye for discriminating between point sources of light is 26 seconds. That is, when light rays from two separate points strike the eye with an angle of at least 26 seconds between them, they can usually be recognized as two points instead of one. This means that a person with maximal acuity looking at two bright pinpoint spots of light 10 meters away can barely distinguish the spots as separate entities when they are 1 millimeter apart.

The fovea is less than one half a millimeter in diameter, which means that maximum visual acuity occurs in only 3 degrees of the visual field. Outside this foveal area the visual acuity is reduced five- to tenfold, and it becomes progressively poorer as the periphery is approached. This is caused by the connection of many rods and cones to the same nerve fiber, as will be discussed in Chapter 60.

Clinical Method for Stating Visual Acuity. Usually the test chart for testing eyes is placed 20 feet away from the tested person, and if the person can see the letters of the size that he should be able to see at 20 feet, he is said to have 20/20 vision, that is, normal vision. If he can see only letters that he should be able to see at 200 feet, he is said to have 20/200 vision. In other words, the clinical method for expressing visual acuity is to use a mathematical fraction that expresses the ratio of two distances, which is also the ratio of one's visual acuity to that of the normal person.

DETERMINATION OF DISTANCE OF AN OBJECT FROM THE EYE—DEPTH PERCEPTION

The visual apparatus normally perceives distance by three major means. This phenomenon is known as *depth perception*. These means are (1) size of the image of known objects on the retina, (2) the phenomenon of moving parallax, and (3) the phenomenon of stereopsis.

Determination of Distance by Sizes of Retinal Images of Known Objects. If one knows that a man whom he is viewing is 6 feet tall, he can determine how far away the man is simply by the size of the man's image on his retina. He does not consciously think about the size, but his brain has learned to calculate automatically from image sizes the distances of objects when the dimensions are known.

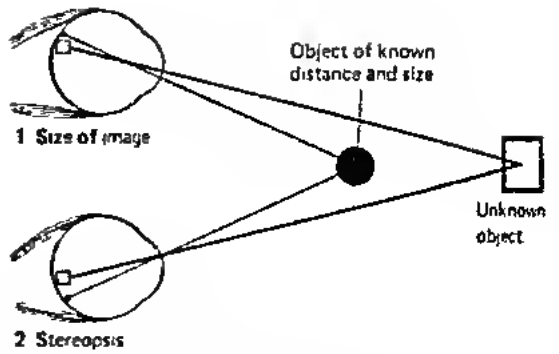


Figure 58–17. Perception of distance (1) by the size of the image on the retina, and (2) as a result of stereopsis.

Determination of Distance by Moving Parallax. Another important means by which the eyes determine distance is that of moving parallax. If a person looks off into the distance with his eyes completely still, he perceives no moving parallax but, when he moves his head to one side or the other, the images of objects close to him move rapidly across his retinae while the images of distant objects remain rather stationary. For instance, if he moves his head 1 inch and an object is only 1 inch in front of his eye, the image moves almost all the way across his retinae, whereas the image of an object 200 feet away from his eyes does not move perceptibly. Thus, by this mechanism of moving parallax, one can tell the *relative distances* of different objects even though only one eye is used.

Determination of Distance by Stereopsis—Binocular Vision. Another method by which one perceives parallax is that of binocular vision. Because one eye is a little more than 2 inches to one side of the other eye, the images on the two retinae are different one from the other—that is, an object that is 1 inch in front of the bridge of the nose forms an image on the temporal portion of the retina of each eye, whereas a small object 20 feet in front of the nose has its image at closely corresponding points in the middle of each retina. This type of parallax is illustrated in Figure 58–17, which shows the images of a black spot and a square actually reversed on the two retinas because they are at different distances in front of the eyes. This gives a type of parallax that is present all the time when both eyes are being used. It is almost entirely this binocular parallax (or stereopsis) that gives a person with two eyes far greater ability to judge relative distances *when objects are nearby* than a person who has only one eye. However, stereopsis is virtually useless for depth perception at distances beyond 200 feet.

OPTICAL INSTRUMENTS

THE OPHTHALMOSCOPE

The ophthalmoscope is an instrument through which an observer can look into another person's eye and see

he retina with clarity. Though the ophthalmoscope appears to be a relatively complicated instrument, its principles are simple. The basic components are illustrated in Figure 58–18 and may be explained as follows.

If a bright spot of light is on the retina of an *emmetropic eye*, light rays from this spot diverge toward the lens system of the eye, and, after passing through the lens system, they are parallel with each other because the retina is located exactly one focal length distance behind the lens. Then, when these parallel rays pass into an emmetropic eye of another person, they focus back again to a point focus on the retina of the second person because his retina is also one focal length distance behind the lens. Therefore, any spot of light on the retina of the observed eye comes to a focal spot on the retina of the observing eye. Likewise, when the bright spot of light is moved to different points on the observed retina, the focal spot on the retina of the observer also moves an equal amount. Thus, if the retina of one person is made to emit light, the image of his retina will be focused on the retina of the observer provided the two eyes are simply looking into each other. These principles, of course, apply only to completely emmetropic eyes.

To make an ophthalmoscope, one need only devise a means for illuminating the retina to be examined. Then, the reflected light from that retina can be seen by the observer simply by putting the two eyes close to each other. To illuminate the retina of the observed eye, an angulated mirror or a segment of a prism is placed in front of the observed eye in such a manner, as illustrated in Figure 58–18, that light from a bulb is reflected into the observed eye. Thus, the retina is illuminated through the pupil, and the observer sees into the subject's pupil by looking over the edge of the mirror or prism, or *through* an appropriately designed prism so that the light will not have to enter the pupil at an angle.

It was noted above that these principles apply only to persons with completely emmetropic eyes. If the refractive power of either eye is abnormal, it is necessary to correct this refractive power in order for the observer to see a sharp image of the observed retina. Therefore, the usual ophthalmoscope has a series of about 20 lenses mounted on a turret so that the turret can be rotated from one lens to another, and the correction for abnormal refractive power of either or both eyes can be made by selecting a lens of appropriate strength. In normal young adults, when the two eyes come close together a natural accommodative reflex occurs that causes an approximate +2 diopters increase in the strength of the lens of each eye. To correct for this, it is necessary that the lens turret be rotated to approximately −4 diopters correction.

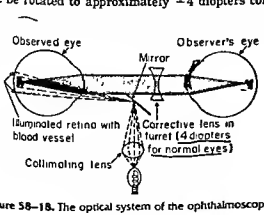


Figure 58–18. The optical system of the ophthalmoscope.

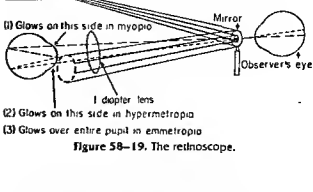


Figure 58–19. The retinoscope.

THE RETINOSCOPE

The retinoscope, illustrated in Figure 58–19, is an instrument that can be used to determine the refractive power of an eye even though the subject cannot converse with the observer. Such a procedure is valuable for fitting glasses to an infant.

To use the retinoscope, one places a bright spot of light behind and to one side of the observed eye, and the observer stands 1 meter away, looking through a hole in the middle of a small round mirror. The observer then rotates this mirror from side to side, making a reflected beam of light travel across the pupil of the observed eye while the subject keeps his gaze intently on the observer's eye. If the observed eye is exactly focused on the observer's eye, when the edge of this beam of light first enters the pupil the entire pupil suddenly glows red. If the eye has abnormal refractive power, the red glow appears either on the side of the pupil into which the light edge first shines or on the opposite side of the pupil—one or the other. *In hypermetropia the first glow appears on the side of the pupil from which the light beam is being moved. In myopia the first glow appears on the opposite side of the pupil.* The cause of this difference is that the lens of the hyperopic eye is too weak and that of the myopic eye too strong to focus the light from the edge of the pupil on the retina exactly in line with the observer's eye. For a fuller understanding of this effect, however, the student is referred to texts on physiologic optics.

One can fit glasses to a patient by placing selected lenses in front of the observed eye one at a time until the glow suddenly covers the pupil over its entire extent rather than spreading from one side of the pupil to the other. However, it should be noted that in retinoscopy one tests an eye that is focused on the observer's eye at 1 meter's distance. This must be taken into consideration in prescribing glasses; 1 diopter strength must be subtracted for far vision.

REFERENCES

Allen, E W Essentials of Ophthalmic Optics. New York, Oxford University Press, 1979

Campell, C J, *et al*. Physiological Optics. Hagerstown, Md., Harper & Row, 1974

Collins, R., and Van der Werff, T J Mathematical Models of the Dynamics of the Human Eye New York, Springer-Verlag, 1980

Davson, H. The Physiology of the Eye, 3rd Ed New York, Academic Press, 1972

Davson, H, and Graham, L. T, Jr.: The Eye. Vols. 1-6 New York, Academic Press, 1969-1974.

Dick, G L: Studies in Ocular Anatomy and Physiology Kensington, N S. W., New South Wales University Press, 1976

Duncan, G., and Jacob, T J Calcium and the physiology of cataract *Ciba Found Symp*, *106* 132, 1984
Fischbarg, J., and Lim, J J Fluid and electrolyte transports across corneal endothelium *Curr Top Eye Res*, *4* 201, 1984
Hartstein J Basics of Contact Lenses, 3rd Ed San Francisco, American Academy of Ophthalmology, 1979
Jaffe N S Cataract Surgery and Its Complications St Louis C V Mosby, 1983
Kayner, H S, and Dusky, L Total Vision New York, A & W Publishers 1980
Kolder, H E J W (ed.) Cataracts. Boston, Little, Brown 1978
Kuszak, J R, *et al* Sutures of the crystalline lens A review *Scan Electron Microsc*, *(Pt 3)* 1369, 1984
Lerman, S Radiant Energy and the Eye New York, The Macmillan Co, 1979
Lesperance Ophthalmic Lasers Photocoagulation, Photoradiation and Surgery St Louis, C V Mosby, 1983
Miller, D Ophthalmology: The Essentials Boston, Houghton Mifflin, 1979
Morgan, M W The Optics of Ophthalmic Lenses Chicago, Professional Press, 1978
Moses R A Adler's Physiology of the Eye; Clinical Application 7th Ed St. Louis, C V Mosby, 1981
Polyak, S The Vertebrate Visual System Chicago, University of Chicago Press, 1957
Records, R E Physiology of the Human Eye and Visual System Hagerstown, Md, Harper & Row, 1979
Regan, D, *et al* The visual perception of motion in depth. *Sci Am* *241(1)* 136, 1979
Roth, H W, and Roth-Wittig M: Contact Lenses Hagerstown, Md Harper & Row 1980
Ruben, M Understanding Contact Lenses (and the Correction of the Abnormal Eye) London, Heinemann Health Books, 1975
Safir A (ed) Refraction and Clinical Optics Hagerstown, Md, Harper & Row, 1980
Sloane, A E (ed) Manual of Refraction Boston, Little, Brown, 1979
Tostes F M Accommodation function of the human eye *Physiol Rev* *52* 828 1972
Van Heyningen, R. What happens to the human lens in cataract. *Sci Am*, *233(6)* 70, 1975
Whitnall, S E The Anatomy of the Human Orbit and Accessory Organs of Vision Huntington, N Y, R E Kreiger Publishing Co, 1979
Whitteridge, D Binocular vision and cortical function *Proc R. Soc Med* *65* 947 1972
Witkovsky P Peripheral mechanisms of vision *Annu Rev Physiol* *33* 257, 1971
Yellott, J I, Jr, *et al* The beginnings of visual perception The retinal image and its initial encoding *In* Dorian-Smith, I (ed.) Handbook of Physiology Sec 1, Vol III Bethesda, American Physiological Society 1984 p 257

59

The Eye: II. Receptor and Neural Function of the Retina

Cones – [illegible]
Rods – [illegible] In Dark

The retina is the light-sensitive portion of the eye, containing the cones which are responsible for color vision and the rods which are mainly responsible for vision in the dark. When the rods and cones are excited, signals are transmitted through successive neurons in the retina itself and finally into the optic nerve fibers and cerebral cortex. The purpose of the present chapter is to explain specifically the mechanisms by which the rods and cones detect both white and colored light.

ANATOMY AND FUNCTION OF THE STRUCTURAL ELEMENTS OF THE RETINA

The Layers of the Retina. Figure 59–1 shows the functional components of the retina arranged in layers from the outside to the inside as follows (1) pigment layer, (2) layer of rods and cones projecting into the pigment, (3) outer limiting membrane, (4) outer nuclear layer, (5) outer plexiform layer, (6) inner nuclear layer, (7) inner plexiform layer, (8) ganglionic layer, (9) layer of optic nerve fibers, and (10) inner limiting membrane.

After light passes through the lens system of the eye and then through the vitreous humor, it enters the retina from the inside (see Figure 59–1), that is, it passes through the ganglion cells, the plexiform layer, the nuclear layer, and the limiting membranes before it finally reaches the layer of rods and cones located all the way on the outer side of the retina. This distance is a thickness of several hundred microns; visual acuity is obviously decreased by this passage through such nonhomogeneous tissue. However, in the central region of the retina, as will be discussed below, the initial layers are pulled aside to prevent this loss of acuity.

The Foveal Region of the Retina and Its Importance in Acute Vision. A minute area in the center of the retina, illustrated in Figure 59–2, called the *macula* and occupying a total area of less than 1 square millimeter, is especially capable of acute and detailed vision. The central portion of the macula, only 0.4 millimeter in diameter, is called the *fovea;* this area is composed entirely of cones, and the cones have a special structure that aids their detection of detail in the visual image,

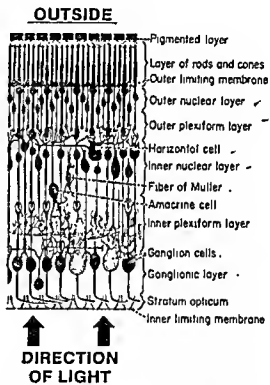


Figure 59–1. Plan of the retinal neurons. (Modified from Polyak: The Retina. Chicago, University of Chicago Press.)

especially a long slender body in contradiction to much larger cones located further peripherally in the retina. Also, in this region the blood vessels, the ganglion cells, the inner nuclear layer of cells, and the plexiform layers are all displaced to one side rather than resting directly on top of the cones. This allows light to pass unimpeded to the cones rather than through several layers of retina, which aids immensely in the acuity of visual perception.

The Rods and Cones. Figure 59–3 is a diagrammatic representation of a photoreceptor (either a rod or a cone) though the cones are distinguished by having a conical upper end (the outer segment), as shown in Figure 59–4. In general, the rods are narrower and longer than the cones, but this is not always the case. In the peripheral portions of the retina the rods are 2 to 5 microns in diameter whereas the cones are 5 to 8 microns

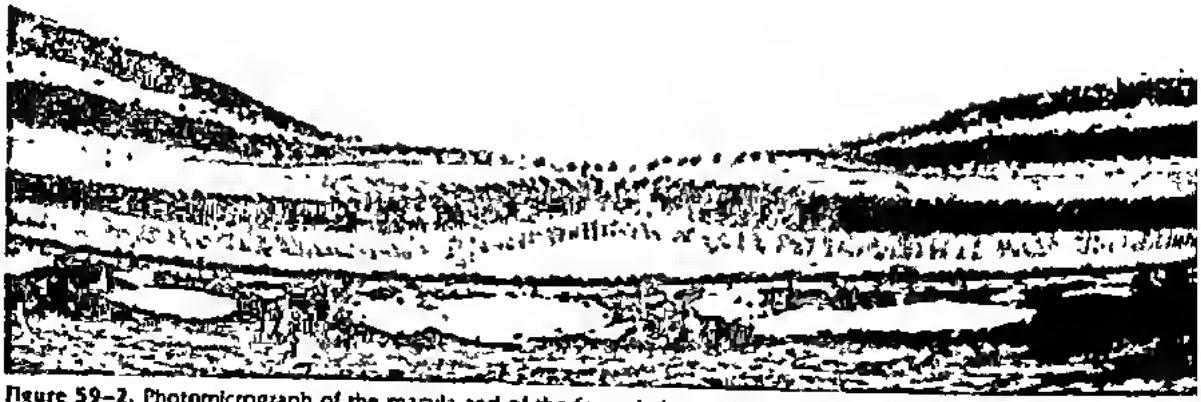


Figure 59–2. Photomicrograph of the macula and of the fovea in its center. Note that the inner layers of the retina are p to the side to decrease the interference with light transmission. (from Bloom and Fawcett: A Textbook of Histology, 10 Philadelphia, W. B. Saunders Company, 1975; courtesy of H. Mizoguchi.)

in diameter; in the central part of the retina, in the fovea, the cones have a diameter of only 1.5 microns.

To the right in Figure 59–3 are labeled the four major functional segments of either a rod or a cone: (1) the *outer segment*, (2) the *inner segment*, (3) the *nucleus*, and (4) the *synaptic body*. In the outer segment the light-sensitive photochemical is found. In the case of the rods, this is *rhodopsin*, and in the cones it is one of several photochemicals collectively called *iodopsin*, which are almost exactly the same as rhodopsin except for differences in spectral sensitivity.

Note both in Figures 59–3 and 59–4 the large numbers of discs in both the rods and the cones. In the cones,

Rods – Rhodopsin
Cones – Iodopsin

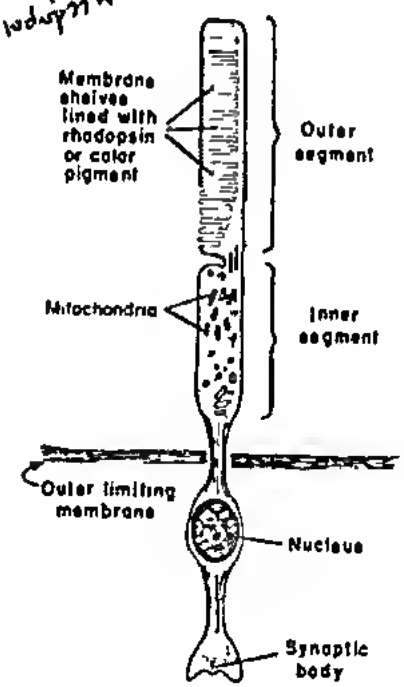


Figure 59–3. Schematic drawing of the functional parts of the and cones.

each of the discs is actually an infolded shelf of membrane, in the rods this is also true near the ba the rod. However, toward the tip of the rod the separate from the membrane and are flat sacs l totally inside the cell. There are as many as 1000 in each rod or cone. At the tips of the rods, the discs continually degenerating, but new ones are also b formed at a rate of about 50 to 100 per day at the of the outer segment. In the cones, similar degenera and replacement occurs, but it occurs along the en extent of the outer segment.

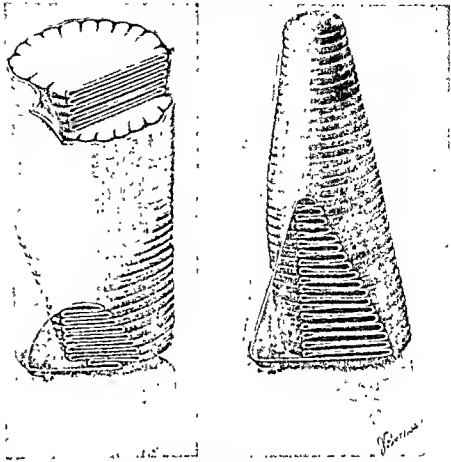
Both rhodopsin and iodopsin are conjugated prote These are incorporated into the membranes of the d in the form of transmembrane proteins. The concen tions of these photosensitive pigments in the disc ar great that they constitute approximately 40 per cen the entire mass of the outer segment.

The inner segment contains the usual cytoplasm the cell with the usual cytoplasmic organelles. Part larly important are the mitochondria, for we shall later that the mitochondria in this segment play important role in providing most of the energy function of the photoreceptors.

The synaptic body is the portion of the rod and c that connects with the subsequent neuronal cells, horizontal and bipolar cells, that represent the n stages in the vision chain.

The Pigment Layer of the Retina. The black p ment *melanin* in the pigment layer, and still m melanin in the choroid, prevent light reflection throu out the globe of the eyeball, this is extremely importa for clear vision. This pigment performs the same fu tion in the eye as does the black coloring inside t bellows of a camera. Without it, light rays would reflected in all directions within the eyeball and wou cause diffuse lighting of the retina rather than t contrast between dark and light spots required for f mation of precise images.

The importance of melanin in the pigment layer a choroid is well illustrated by its absence in *albin* persons hereditarily lacking in melanin pigment in parts of their bodies. When an albino enters a brig area, light that impinges on the retina is reflected in



ure 59–4. Membranous structures of the segments of a rod (left) and a cone (right) ntesy of Dr. Richard Young.)

ections by the white surface of the unpigmented roid so that a single discrete spot of light that would rmally excite only a few rods or cones is reflected rywhere and excites many of the receptors There- , the visual acuity of albinos, even with the best of ical correction, is rarely better than 20/100 to 20/200 'he pigment layer also stores large quantities of *amin A*. This vitamin A is exchanged back and forth ough the membranes of the outer segments of the s and cones, which themselves are embedded in the ment layer. We shall see later that vitamin A is an portant precursor of the photosensitive pigments and t this interchange of vitamin A is very important for ustment of the light sensitivity of the receptors

'he Blood Supply of the Retina—The Retinal terial System and the Choroid. The nutrient blood ply for the inner layers of the retina is derived from central retinal artery, which enters the inside of the along with the optic nerve and then divides to supply entire inner retinal surface. Thus, to a great extent, retina has its own blood supply independent of the er structures of the eye.

However, the outer layer of the retina is adherent to *choroid*, which is a highly vascular tissue between retina and the sclera. The outer layers of the retina, luding the *outer segments of the rods and cones*, pend mainly on diffusion from the choroid vessels for ir nutrition, especially for their oxygen

Retinal Detachment. The neural retina occasionally aches from the pigment epithelium In some in- nces the cause of such detachment is injury to the eball that allows fluid or blood to collect between the ina and the pigment epithelium, but often it is also ised by contracture of fine collagenous fibrils in the vitreous humor, which pull the retina unevenly toward the interior of the globe.

Fortunately, partly because of diffusion across the detachment gap and partly because of the independent blood supply to the retina through the retinal artery, the detached retina can resist degeneration for days and can become functional once again if surgically replaced in its normal relationship with the pigment epithelium But, if not replaced soon, the retina finally does degenerate and is then unable to function even after surgical repair

PHOTOCHEMISTRY OF VISION

Both the rods and cones contain chemicals that decompose on exposure to light and, in the process, excite the nerve fibers leading from the eye. The chemical in the *rods* is called *rhodopsin*, and the light-sensitive chemicals in the *cones* have compositions only slightly different from that of rhodopsin and are collectively called *iodopsin*.

In the present section we will discuss principally the photochemistry of rhodopsin, but we can apply almost exactly the same principles to the photochemistry of the iodopsin of the cones.

THE RHODOPSIN-RETINAL VISUAL CYCLE, AND EXCITATION OF THE RODS

Rhodopsin and Its Decomposition by Light Energy. The outer segment of the rod that projects

into the pigment layer of the retina has a concentration of about 40 per cent of the light-sensitive pigment called *rhodopsin*, or *visual purple*. This substance is a combination of the protein *scotopsin* and the carotenoid pigment *retinal* (also called "retinene"). Furthermore, the retinal is a particular type called 11-*cis* retinal. This *cis* form of the retinal is important because only this form can combine with scotopsin to synthesize rhodopsin.

When light energy is absorbed by rhodopsin, the rhodopsin immediately begins to decompose, as shown at the top of Figure 59–5. The cause of this is photoactivation of electrons in tha retinal portion of the rhodopsin, which leads to an instantaneous change (in the order of trillionths of a second) of the *cis* form of retinal into an all-*trans* form, which still has the same chemical structure as the cis form but has a different physical structure—a straight molecule rather than a curved molecule. Because the three-dimensional orientation of the reactive sites of the all-*trans* retinal no longer fits with that of the reactive sites on the protein scotopsin, it begins to pull away from the scotopsin. The immediate product is *prelumirhodopsin*, which is a partially split combination of the all-*trans* retinal and scotopsin. However, prelumirhodopsin is an extremely unstable compound and decays in nanoseconds to *lumirhodopsin*. This then decays in microseconds to *metarhodopsin I*, then in about a millisecond to *metarhodopsin II*, finally, much more slowly (in seconds) into the completely split products: *scotopsin* and *all-trans retinal*. During the first stages of splitting, the rods are excited and signals are transmitted into the central nervous system, as we shall discuss later.

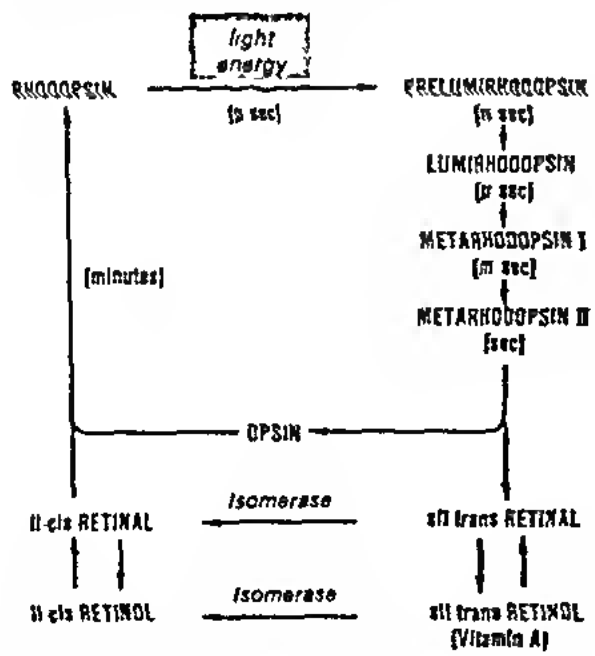


Figure 59–5. Photochemistry of the rhodopsin-retinal vitamin A visual cycle

Reformation of Rhodopsin. The first st reformation of rhodopsin, as shown in Figu 5, is to reconvert the all-*trans* retinal into retinal. This process is catalyzed by the e *retinal isomerase*. Once the 11-*cis* retinal is f it automatically recombines with the scoto reform rhodopsin, an exergonic process means that it gives off energy). The produc dopsin, is a stable compound until its deco tion is again triggered by absorption of lig ergy.

The Role of Vitamin A in the Formati Rhodopsin. Note in Figure 59–5 that the second chemical route by which all-*trans* r can be converted into 11-*cis* retinal. This conversion of the all-*trans* retinal first in *trans retinol*, which is one form of vitamin A the all-*trans* retinol is converted into 11-*cis* r under the influence of the enzyme isomerase finally the 11-*cis* retinol is converted into retinal.

Vitamin A is present both in the cytopl the rods and in the pigment layer of the ret well Therefore, vitamin A is normally a available to form new retinal when neede the other hand, when there is excess retinal retina, the excess is converted back into vi A, thus reducing the amount of light-se pigment in the retina. We shall see later th interconversion between retinal and vitami especially important in long-term adaptat the retina to different light intensities.

Night Blindness. Night blindness occurs in vitamin A deficiency. The simple reason for this not enough vitamin A is then available to form ad quantities of retinal. Therefore, the amounts of r sin that can be formed in the rods, as well as the a of color-photosensitive chemicals in the cones, depressed This condition is called night blind cause the amount of light available at night is t little to permit adequate vision, though in daylig rods and cones can still be excited despite their re in photochemical substances.

For night blindness to occur, a person usually remain on a vitamin A–deficient diet for months, because large quantities of vitamin A are no stored in the liver However, once night blind develop, it can sometimes be completely cured than an hour by intravenous injection of vitamin

Excitation of the Rod When Rhodopsin Decomposes

Generation of the Rod Receptor Pote Generation of the rod receptor potential is e different from the generation of receptor pote in almost all other sensory receptors Th excitation of the rod causes *increased negati* the membrane potential, which is a state of *h polarization*, rather than decreased nega which is the process of "depolarization" t characteristic of almost all other sensory rece

But, how does the breakdown of rhodopsin

hyperpolarization? The answer to this is that *when rhodopsin decomposes, it decreases the membrane conductance for sodium ions in the outer segment of the rod.* And this causes hyperpolarization of the entire rod membrane in the following way:

Figure 59–6 illustrates movement of sodium ions in a complete electrical circuit through the inner and outer segments of the rod. The inner segment continually pumps sodium from inside the rod to the outside, thereby creating a negative potential on the inside of the entire cell. However, the membrane of the outer segment, in the *dark* state, is very leaky to sodium. Therefore, sodium continually leaks back to the inside of the rod and thereby neutralizes much of the negativity on the inside of the entire cell. Thus, under normal conditions, when the rod is not excited there is a reduced amount of electronegativity inside the membrane of the rod, normally about −30 millivolts.

When the rhodopsin in the outer segment of the rod is exposed to light and begins to decompose, however, this *decreases* the conductance of sodium to the interior of the rod even though sodium ions continue to be pumped out. Thus, more sodium ions now leave the rod then leak back in. Because these are positive ions, their loss from inside the rod creates increased negativity inside the membrane; and the greater the amount of light energy striking the rod, the greater the electronegativity—that is, the greater the degree of *hyperpolarization.* At maximum light intensity, the membrane potential approaches −90 millivolts, which is the equilibrium potential for potassium ions across the membrane.

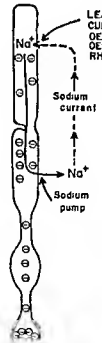


Figure 59–6. Theoretical basis for the generation of a hyperpolarization receptor potential caused by rhodopsin decomposition.

Duration of the Receptor Potential and Logarithmic Relationship of the Receptor Potential to Light Intensity. When a sudden pulse of light strikes the retina, the transient hyperpolarization that occurs—that is, the receptor potential that occurs—lasts for about one twentieth to one half of a second, depending on the intensity of the light and other factors. Therefore, a visual image impinged on the retina for only a millionth of a second nevertheless can still cause the sensation of seeing the image sometimes for up to half a second.

Another characteristic of the receptor potential is that it is approximately proportional to the logarithm of the light intensity. This is exceedingly important, because it allows the eye to discriminate light intensities through a range many thousand times as great as would be possible otherwise.

The Mechanism by Which Rhodopsin Decomposition Decreases Membrane Sodium Conductance. Though it is now well known that the decomposition of rhodopsin by light causes decreased outer segment membrane conductance for sodium ions, the basic mechanism of this is still mainly theoretical. Two important theories are the following:

(1) The discs inside the rods are actually large, flat vesicles, and inside these is a high concentration of calcium ions. It has been postulated that when the transmembrane rhodopsin molecule becomes activated by light, the breaking away of the retinal molecule from the scotopsin molecule opens a calcium channel through this scotopsin. Therefore, quantities of calcium ions discharge instantly from the vesicles into the cytoplasm of the outer segment. These calcium ions in turn diffuse to the cell membrane and bind with the inner surfaces of the membrane's sodium channels. This blocks the sodium channels and decreases the sodium conductance, thus leading to hyperpolarization.

(2) Another theory is that light activation of the rhodopsin molecule triggers a sequence of chemical changes that causes closure of the sodium pores in the following way: (a) Splitting of the rhodopsin activates a phosphodiesterase. (b) The phosphodiesterase causes the breakdown of cyclic GMP. (c) The cyclic GMP normally functions in conjunction with a protein kinase to keep the sodium channels open. However, when light activation diminishes the quantity of cyclic GMP, this now makes it impossible to keep the sodium channels fully opened. Therefore, sodium conductance decreases.

Thus, it is not known exactly how light activation of rhodopsin decreases sodium conductance, but it is this decrease in sodium conductance that leads to the hyperpolarization, which is the receptor potential.

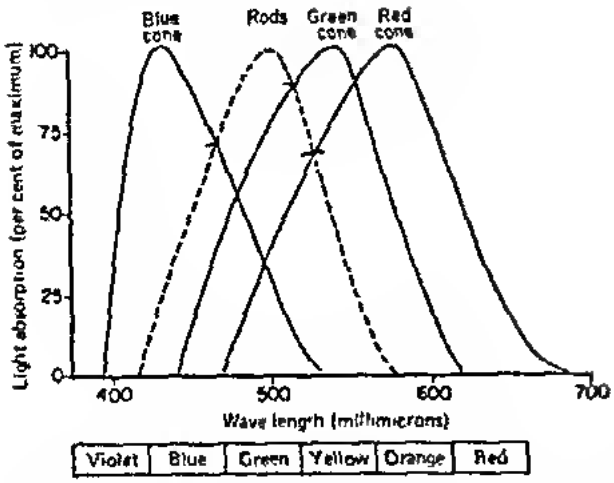


Figure 59–7. Light absorption by the respective pigments of the three color-receptive cones of the human retina. (Drawn from curves recorded by Marks, Dobelle, and MacNichol, Jr. Science, 143 (181) 1964, and by Brown and Wald Science, 144 45, 1964 Copyright 1964 by the American Association for the Advancement of Science.)

PHOTOCHEMISTRY OF COLOR VISION BY THE CONES

It was pointed out at the outset of this discussion that the photochemicals in the cones have almost exactly the same chemical composition as that of rhodopsin in the rods The only difference is that the protein portions, the opsins, called *photopsins* in the cones, are different from the scotopsin of the rods. The retinal portions are exactly the same in the cones as in the rods The color-sensitive pigments of the cones, therefore, are combinations of retinal and photopsins.

In the discussion of color vision later in the chapter, it will become evident that three different types of photochemicals are present in different cones, thus making these cones selectively sensitive to the different colors of blue, green, and red. These photochemicals are called respectively *blue-sensitive pigment, green-sensitive pigment,* and *red-sensitive pigment.* The absorption characteristics of the pigments in the three types of cones show peak absorbancies at light wavelengths respectively, of 445, 535, and 570 millimicrons These are also the wavelengths for peak light sensitivity for each type of cone, which begins to explain how the retina differentiates the colors The approximate absorption curves for these three pigments are shown in Figure 59–7. Also shown is the absorption curve for the rhodopsin of the rods, having a peak at 505 millimicrons.

AUTOMATIC REGULATION OF RETINAL SENSITIVITY—DARK AND LIGHT ADAPTATION

Relationship of Sensitivity to Pigment Concentration. The sensitivity of rods is approximately proportional to the antilogarithm of the rhodopsin concentration, and it is assumed th this relationship also holds true in the con Therefore, the sensitivity of the rods and cones c be altered up or down tremendously by only slig changes in concentrations of the photosensit chemicals.

Light and Dark Adaptation. If a person l been in bright light for a long time, large prop tions of the photochemicals in both the rods s cones have been reduced to retinal and opsi Furthermore, most of the retinal of both the r and cones has also been converted into vitamin Because of these two effects, the concentration the photosensitive chemicals are considerably duced, and the sensitivity of the eye to light even more reduced. This is called *light adaptat*

On the other hand, if the person remains darkness for a long time, essentially all the reti and opsins in the rods and cones become conver into light-sensitive pigments. Furthermore, la amounts of vitamin A are converted into reti which is then changed into additional light-se tive pigments, the final limit being determined the amount of opsins in the rods and cones cause of these two effects, the visual recep gradually become so sensitive that even the nutest amount of light causes excitation. Thi called *dark adaptation.*

Figure 59–8 illustrates the course of dark aptation when a person is exposed to total darkn after having been exposed to bright light for eral hours. Note that sensitivity of the retin very low on first entering the darkness, but wit 1 minute the sensitivity has increased tenfol that is, the retina can respond to light of one te the previously required intensity. At the end o minutes the sensitivity has increased about 60 fold, and at the end of 40 minutes it has incre about 25,000-fold.

The resulting curve of Figure 59–8 is called

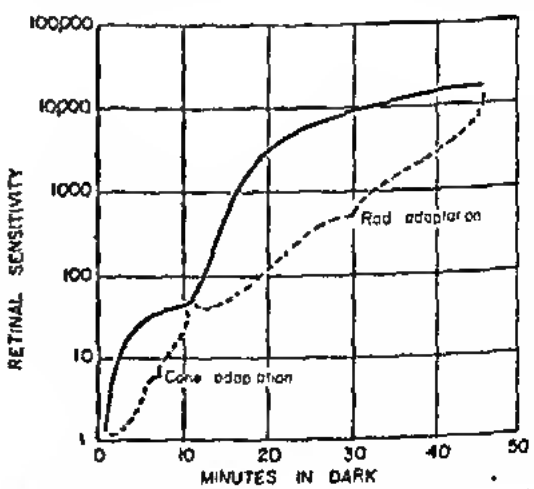


Figure 59–8. Dark adaptation, illustrating the relationsh cone adaptation to rod adaptation.

ork adaptation curve. Note, however, the inflection in the curve. The early portion of the curve is by adaptation of the cones, for these adapt uch more rapidly than the rods because of a basic ifference in the rate at which they resynthesize heir photosensitive pigments. On the other hand, cones do not achieve anywhere near the same egree of sensitivity as the rods. Therefore, despite apid adaptation by the cones, they cease adapting ufter only a few minutes, whereas the slowly adapting rods continue to adapt for many minutes and even hours, their sensitivity increasing tremendously. However, a large share of the greater sensitivity of the rods is also caused by convergence of as many as 100 or more rods onto a single cell in the retina; these rods summate to increase their sensitivity, as will be discussed later in the chapter.

Other Mechanisms of Light and Dark Adaptation. In addition to adaptation caused by changes in concentrations of rhodopsin or color photochemicals, the eye has two other mechanisms for light and dark adaptation. The first of these is the *change in pupillary size*, which was discussed in the previous chapter. This can cause a degree of adaptation of approximately thirtyfold because of changes in the amount of light allowed through the pupillary opening.

The other mechanism is *neural adaptation*, involving the neurons in the successive stages of the visual chain in the retina itself. That is, when the light intensity first increases, the intensities of the signals transmitted by the bipolar cells, the horizontal cells, the amacrine cells, and the ganglion cells are all very intense. However, the intensities of these signals all decrease rapidly. Although the degree of this adaptation is only a fewfold rather than the many thousand-fold that occurs during adaptation of the photochemical system, this neural adaptation occurs in a fraction of a second, in contrast *to the many minutes* required for *full adaptation by the* photochemicals.

Value of Light and Dark Adaptation in Vision. Between the limits of maximal dark adaptation and maximal light adaptation, the eye can change its sensitivity to light by as much as 500,000 to 1,000,000 times, the sensitivity automatically adjusting to changes in illumination.

Since the registration of images by the retina requires detection of both dark and light spots in the image, it is essential that the sensitivity of the retina always be adjusted so that the receptors respond to the lighter areas but not to the darker areas. An example of maladjustment of the retina occurs when a person leaves a movie theater and enters the bright sunlight, for even the dark spots in the images then seem exceedingly bright, and, as a consequence, the entire visual image is bleached, having little contrast between its different parts. Obviously, this is poor vision, and it remains poor until the retina has adapted sufficiently that the darker spots of the image no longer stimulate the receptors excessively.

Conversely, when a person enters darkness, the sensitivity of the retina is usually so slight that even the light spots in the image cannot excite the retina. But, after dark adaptation, the light spots begin to register. As an example of the extremes of light and dark adaptation, the light intensity of the sun is approximately 30,000 times that of the moon; yet the eye can function well both in bright sunlight and in bright moonlight.

Fusion of Flickering Lights by the Retina

In a flickering light, the light intensity alternately increases and decreases rapidly. An instantaneous flash of light excites the visual receptors for as long as 0.05 to 0.5 second, and because of the *persistence* of excitation, rapidly successive flashes of light become *fused* together to give the appearance of being continuous. This effect is well known when one observes motion pictures or television. The images on the motion picture screen are flashed at a rate of 24 frames per second, while those of the television screen are flashed at a rate of 60 frames per second. As a result, the images fuse together, and continuous motion is observed.

The frequency at which flicker fusion occurs, called the *critical frequency for fusion*, varies with the light intensity. Figure 59–9 illustrates the effect of intensity of illumination on the critical fusion frequency. At a low intensity, fusion results even when the rate of flicker is as low as 2 to 6 per second. However, in bright illumination, the critical frequency for fusion rises to as great as 60 flashes per second. This difference results at least partly from the fact that the cones, which operate mainly at high levels of illumination, can detect much more rapid alterations in illumination than can the rods, which are the important receptors in dim light.

COLOR VISION

From the preceding sections, we know that different cones are sensitive to different colors of light. The present section is a discussion of the mechanisms by which the retina detects the different gradations of color in the visual spectrum.

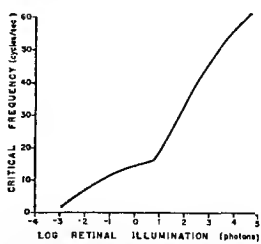


Figure 59–9. Relationship of intensity of illumination to the critical frequency for fusion.

THE TRI-COLOR MECHANISM OF COLOR PERCEPTION

Many different theories have been proposed to explain the phenomenon of color vision, but they are all based on the well-known observation that the human eye can detect almost all gradations of colors when red, green, and blue monochromatic lights are appropriately mixed in different combinations

The first important theory of color vision was that of Young, which was later expanded and given a more experimental basis by Helmholtz Therefore, the theory is known as the *Young-Helmholtz theory*. According to this theory, there are three different types of cones, each of which responds maximally to a different color

As time has gone by, the Young-Helmholtz theory has been expanded, and more details have been worked out It now is generally accepted as the mechanism of color vision

Spectral Sensitivities of the Three Types of Cones. On the basis of color vision tests, tha spectral sensitivities of the three different types of cones in human beings have been proved to be essentially the same as the light absorption curves for the three types of pigment found in the respective cones These were illustrated in Figure 59–7 and are also shown in Figure 59–10 These curves can readily explain almost all the phenomena of color vision

Interpretation of Color in the Nervous System. Referring to Figure 59–10, one can see that an orange monochromatic light with a wavelength of 580 millimicrons stimulates the red cones to a stimulus value of approximately 99 (99 per cent of the peak stimulation at optimum wavelength), whereas it stimulates the green cones to a stimulus value of approximately 42 and the blue rones not at all Thus, the ratios of stimulation of the three different types of cones in this instance are 99 42.0. The nervous system interprets this set of ratios as the sensation of orange On the other hand, a monochromatic blue light with a wavelengt 450 millimicrons stimulates the red cones stimulus value of 0, the green cones to a val 0, and the blue cones to a value of 97. This s ratios—0:0.97—is interpreted by the nervous tem as blue. Likewise, ratios of 83.83.0 are i preted as yellow and 31:67:36, as green.

This scheme also shows how it is possible f person to perceive a sensation of yellow wh red light and a green light are shone into the at the same time, for this stimulates the red green cones approximately equally, which giv sensation of yellow even though no wavelengt light corresponding to yellow is present.

Perception of White Light. Approxima equal stimulation of all the red, green, and b rones gives one the sensation of seeing white there is no wavelength of light corresponding white, instead, white is a combination of all wavelengths of the spectrum. Furthermore, sensation of white can be achieved by stimulat the retina with a proper combination of only th chosen colors that stimulate the respective ty of cones approximately equally.

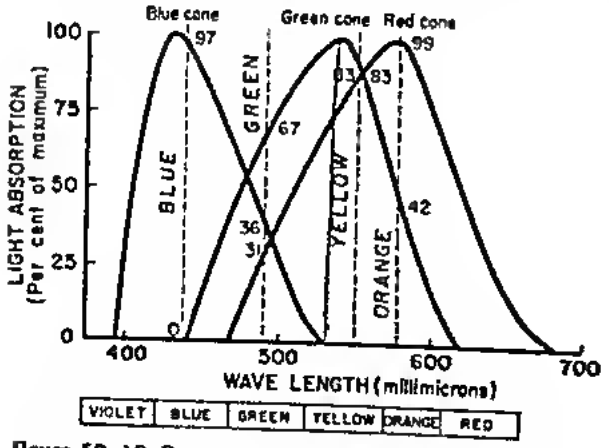


Figure 59–10. Demonstration of the degree of stimulation of the different color-sensitive cones by monochromatic lights of four separate colors. blue, green, yellow, and orange

COLOR BLINDNESS

Red-Green Color Blindness. When a single gr of color receptive cones is missing from the eye, person is unable to distinguish some colors from othe As can be observed by studying Figure 59–10, if the cones are missing, light of 525 to 675 millimicro wavelength can stimulate only the green sensit cones, so that the *ratio* of stimulation of the differ cones does not change as the color changes from gre all the way through the red spectrum. Therefore, with this wavelength range, all colors appear to be the sa to this "color blind" person

On the other hand, if the green-sensitive cones a missing, the colors in the range from green to red c stimulate only the red-sensitive cones, and the pers again perceives only one color within these limi Therefore, when either the red or green types of con are lacking, the person is said to be "red-green" col blind However, when one or more types of cones s abnormal but still function partially, a person has "col weakness," instead of color blindness.

The person with loss of red cones is called a *protanop* his overall visual spectrum is noticeably shortened the long wavelength end because of lack of the red cone The color blind person who lacks green cones is called *deuteranope*; this person has a perfectly normal visu spectral width because the absent green cones operat in the middle of the spectrum where red or blue cone also operate.

Blue Weakness. Occasionally, a person has "blu weakness," which results from diminished or absen blue receptors. If we observe Figure 59–10 still again we can see that the blue cones are sensitive to a spectra renge almost entirely different from that of both the re and green cones. Therefore, if the blue receptors are completely absent, the person has a greater preponder ance of green, yellow, orange, and red in the visual spectrum than of blue, thus producing this rarely observed type of color weakness or blindness.

arranged with a confusion of spots of several different colors In the top chart, the normal person reads "74," while the red-green color blind person reads "21 " In the bottom chart, the normal person reads "42," while the red blind protanope reads "2," and the green blind deuteranope reads "4 "

If one will study these charts while at the same time observing the spectral sensitivity curves of the different cones in Figure 59–10, it can be readily understood how excessive emphasis can be placed on spots of certain colors by color blind persons in comparison with normal persons

Genetics of Color Blindness. Color blindness is sex-linked and results from absence of appropriate color genes in the X chromosomes This lack of color genes is a recessive trait, therefore, color blindness will not appear as long as another X chromosome carries the genes necessary for development of the respective color-receptive cones

Because the male human being has only one X chromosome, all three color genes must be present in this single chromosome if he is not to be color blind In approximately 1 of every 50 times, the X chromosome lacks the red gene, in approximately 1 of every 16, it lacks the green gene; and, rarely, it lacks the blue gene. This means, therefore, that 2 per cent of all men are red color blind (protanopes) and 6 per cent are green color blind (deuteranopes), making a total of approximately 8 per cent who are red-green color blind Because a female has two X chromosomes, red-green color blindness is a rare abnormality in the female

NEURAL FUNCTION OF THE RETINA

NEURAL ORGANIZATION OF THE RETINA

The detailed anatomy of the retina was illustrated in Figure 59–1, Figure 59–12 illustrates the basic essentials of the retina's neural connections, to the left is the general organization of the neural

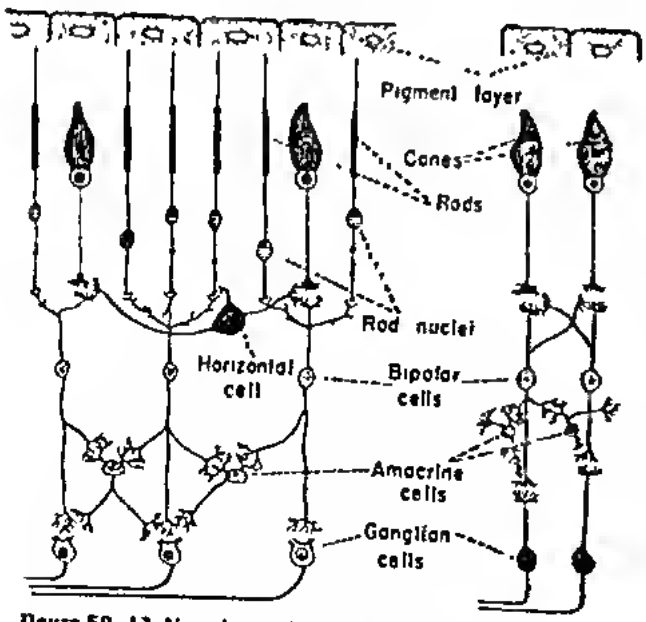


Figure 59–12. Neural organization of the retina, peripheral area to the left, foveal area to the right

elements in a peripheral retinal area and to t right in the foveal area. Note that in the periphe region both rods and cones converge on *bipol cells* which in turn converge on *ganglion cells* the fovea, where only cones exist, there is lit convergence; instead, the cones are represented approximately equal numbers of bipolar and ga glion cells

Not emphasized in the figure is the fact that both the outer and inner plexiform layers (whe the neurons synapse with each other) are ma lateral connections between the different neur elements. Many of these lateral connections a collateral branches of the dendrites and axons the bipolar and ganglion cells. But, in additio two special types of cells are present in the inn nuclear layer that transmits signals laterally: (1 the *horizontal cells*, and (2) the *amacrine cells*

Each retina contains about 125 million rods ar 5.5 million cones; yet, as counted with the ligh microscope, only 1,000,000 optic nerve fibers lea from the retina to the brain. Thus, an average about 125 rods and 5 cones converge on each opti nerve fiber. However, there are major difference between the peripheral retina and the centra retina, for nearer the fovea fewer and fewer rod and cones converge on each optic fiber, and th rods and cones both become slenderer. These tw effects progressively increase the acuity of visio toward the central retina. In the very centra portion, in the *fovea*, there are no rods at all Also the number of optic nerve fibers leading from thi part of the retina is almost equal to the numbe of cones, as shown to the right in Figure 59–12 This mainly explains the high degree of visual acuity in the central portion of the retina in comparison with the very poor acuity in the peripheral portions

Another difference between the peripheral and central portions of the retina is the considerably greater sensitivity of the peripheral retina to weak light This results partly from the fact that as many as 300 rods converge on the same optic nerve fiber in the most peripheral portions of the retina; the signals from the rods summate to give intense stimulation of the peripheral ganglion cells. But, in addition, rods are more sensitive to weak light than are the cones because of intrinsic differences in the receptors themselves.

STIMULATION OF THE RODS AND CONES

Neither the rods nor the cones generate action potentials Instead, the hyperpolarization receptor potentials generated in the outer segments of the rods and cones are transmitted through the bodies of these receptors to the *synaptic bodies* at their other ends by direct conduction of the electrical voltage itself, this is called *electrotonic conduction.* Then, at the synaptic body, the receptor potential controls the release of a transmitter substance the

hemical nature of which is still not known. However, in the dark state, when the receptor is depolarized, the rate of transmitter release is greatest, and its release decreases markedly as the receptor becomes hyperpolarized in the full light condition. The transmitter in turn induces signals in the successive neurons, the bipolar and horizontal cells. The signals in both these cells are also transmitted by electrotonic conduction and not by action potentials.

STIMULATION OF THE BIPOLAR CELLS

The most direct pathway from the rods and cones to the ganglion cells is through the bipolar cells. However, there are two different types of bipolar cells, the *depolarizing bipolar cell* and the *hyperpolarizing bipolar cell*.

The depolarizing bipolar cell is stimulated by the rods and cones when they are exposed to light. The mechanism of this is peculiar in the following way: The transmitter substance from the rods and cones inhibits the depolarizing bipolar cell. Therefore, in the dark, when the rods and cones are secreting large quantities of the transmitter substance, the depolarizing bipolar cell is inhibited. But in the light state, the reduction of transmitter substance from the rods and cones now reduces the inhibition of the bipolar cell, allowing it to become *excited*.

The hyperpolarizing bipolar cell is excited, not inhibited, by the same transmitter from the rods and cones that inhibits the depolarizing cell because its receptors are of the excitatory variety, not the inhibitory variety. Therefore, in the light, exactly opposite events occur in the hyperpolarizing cell to those in the depolarizing cell. That is, the hyperpolarizing cell becomes inhibited.

Because of this difference between the two different types of bipolar cells, both positive signals and negative signals can be transmitted through different bipolar cells from the rods and cones to the amacrine and ganglion cells (Both types of bipolar cells release an excitatory transmitter at their endings that causes excitation of the amacrine and ganglion cells.)

STIMULATION AND FUNCTION OF THE HORIZONTAL CELLS

The horizontal cells, illustrated to the left in Figure 59–12, connect laterally in the outer plexiform layer of the retina. The input to these cells is principally from the rods and cones; in turn, they excite the dendrites of bipolar cells located many microns laterally.

The horizontal cells respond to the transmitter from the rods and cones in the same manner as the depolarizing bipolar cells. That is, they are excited by light. However, they in turn release an inhibitory transmitter that inhibits the laterally displaced bipolar cells with which they connect. Therefore, the horizontal cells represent a *lateral inhibitory pathway* in the retina.

STIMULATION AND FUNCTION OF THE AMACRINE CELLS

The amacrine cells are excited by the bipolar cells and they in turn excite the ganglion cells. Many of their fiber pathways travel laterally in the inner plexiform layer of the retina so that this represents another pathway by which lateral signals can be transmitted, though this time the signals are excitatory in contrast to the inhibitory signals of the horizontal cells.

Perhaps the most important characteristic of the amacrine cells is that when they are stimulated by the bipolar cells, they at first respond very strongly, but the signal dies away to almost nothing in a fraction of a second. Therefore, it is believed that the amacrine cells send strong signals to the brain to indicate sudden changes in light intensity.

Another difference between amacrine cells and the bipolar and horizontal cells is that amacrine cells often transmit action potentials (but can transmit signals by electrotonic conduction as well); the significance of this difference is not known.

EXCITATION OF THE GANGLION CELLS

Spontaneous, Continuous Discharge of the Ganglion Cells. The ganglion cells transmit their signals through the optic nerve fibers to the brain in the form of action potentials These cells, even when unstimulated, transmit continuous nerve impulses at an average rate of about 5 per second. The visual signal is superimposed onto this basic level of ganglion cell stimulation. It can be either an excitatory signal, with the number of impulses increasing to greater than 5 per second, or an inhibitory signal, with the number of nerve impulses decreasing to below 5 per second—often all the way to zero.

Summation at the Ganglion Cells of Signals from the Bipolar Cells, the Horizontal Cells, and the Amacrine Cells. The depolarizing bipolar cells transmit the main direct *excitatory* information from the rods and the cones to the ganglion cells: the hyperpolarizing bipolar cells and the horizontal cells transmit *inhibitory* information from laterally displaced rods and cones to surrounding bipolar cells and then to the ganglion cells; the amacrine cells seem to transmit direct but short-lived transient signals that signal a *change* in the level of illumination of the retina. Thus, each of these types of cells performs a separate function in stimulating the ganglion cells.

Also, there are at least three different types of ganglion cells, as we shall discuss in the following chapter. Some of these transmit very discrete signals from narrow groups of visual receptors Oth-

ers have widespreading dendritic trees that collect visual information from large areas of the retina and therefore transmit general information, such as luminosity or a sudden change in the visual scene, rather than transmitting the specific visual pattern.

DIFFERENT TYPES OF SIGNALS TRANSMITTED BY THE GANGLION CELLS THROUGH THE OPTIC NERVE

Transmission of Signals Depicting Contrasts in the Visual Scene—The Role of Lateral Inhibition

Most of the ganglion cells do not respond to the actual level of illumination of the scene, instead they respond only to contrast borders in the scene. Since it seems that this is the major means by which the form of the scene is transmitted to the brain, let us explain how this process occurs.

When flat light is applied to the entire retina—that is, when all the photoreceptors are stimulated equally by the incident light—the contrast type of ganglion cell is neither stimulated nor inhibited. The reason for this is that the signals transmitted *directly* from the photoreceptors through the depolarizing bipolar cells are excitatory, whereas the signals transmitted *laterally* through the hyperpolarizing bipolar cells and horizontal cells are inhibitory. Thus, the direct excitatory signal through one pathway is likely to be completely neutralized by the inhibitory signals through the lateral pathways. One circuit for this is illustrated in Figure 59–13, which shows three photoreceptors, the central one of these receptors excites a depolarizing bipolar cell. However, the two receptors on either side are connected to the same bipolar cell through inhibitory horizontal cells that neutralize the direct excitatory signal if these receptors are also stimulated by light.

Now, let us examine what happens when a contrast border occurs in the visual scene. Referring again to Figure 59–13, let us assume that the central photoreceptor is stimulated by a bright spot of light while the two lateral receptors are in the dark. The bright spot of light will excite the direct pathway through the bipolar cell. Then, in addition, the fact that the two lateral photoreceptors are in the dark causes the two horizontal cells to be inhibited. In turn, these cells lose their inhibitory effect on the bipolar cell, and this allows still more excitation of the bipolar cell. Thus, when light is everywhere, the excitatory and inhibitory signals to the bipolar cells mainly neutralize each other, but where contrasts occur the signals through the direct and lateral pathways actually accentuate each other.

Thus, the mechanism of lateral inhibition functions in the eye in the same way that it functions in most other sensory systems as well—that is, to provide contrast detection and enhancement.

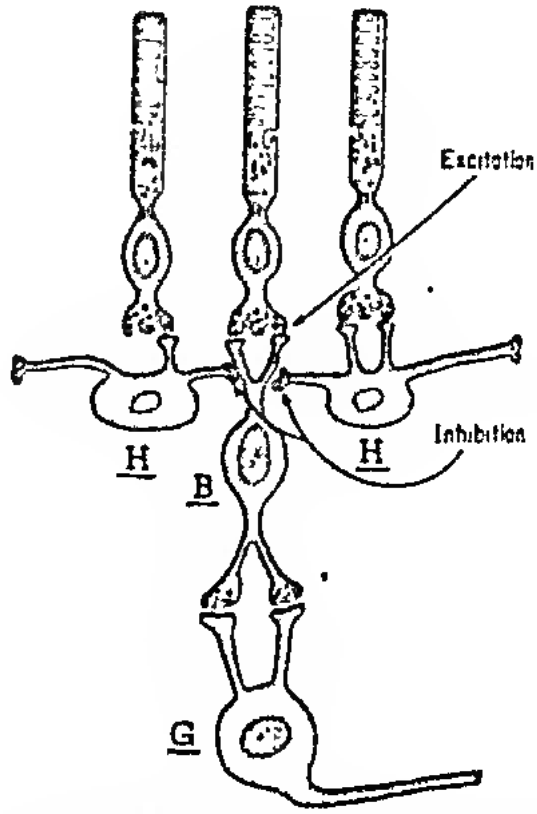


Figure 59–13. Typical arrangement of rods, horizontal cells (H), a bipolar cell (B) and a ganglion cell (G) in the retina, showing excitation at the synapses between the rods and the horizontal cells but inhibition between the horizontal cells and the bipolar cells.

Detection of Instantaneous Changes in Light Intensity—The On-Off Response. Many of the ganglion cells are especially excited by *change* in light intensity. For instance, the upper tracing of Figure 59–14 shows that when the light was first turned on the ganglion cell became strongly excited for a fraction of a second, and then the level of excitation diminished. The bottom tracing shows that a second ganglion cell located in the dark area lateral to the spot of light was markedly inhibited at the same time because of lateral inhibition. Then when the light was turned off exactly the opposite effects occurred. Thus, the responses of these two types of cells are called the "on-off" and the "off-on" responses.

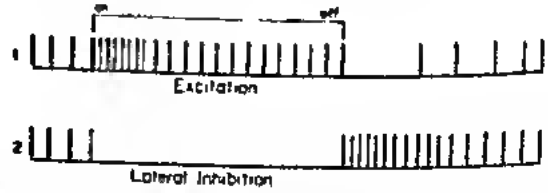


Figure 59–14. Responses of ganglion cells to light in (1) an area excited by a spot of light and (2) an area immediately adjacent to the excited spot, the ganglion cells in this area are inhibited by the mechanism of lateral inhibition. (Modified from Granit, Receptors and Sensory Perception. A Discussion of Aims, Means, and Results of Electrophysiological Research into the Process of Reception. Yale University Press, 1955.)

This ability of the retina to detect and transmit signals related to *change* in light intensity is caused by a rapid phase of "adaptation" of some of the neurons in the visual chain. Since this effect is known to be extremely marked in the amacrine cells, it has been suggested that the amacrine cells are peculiarly adapted to detecting light intensity change.

This capability to detect change in light intensity is especially well developed in the peripheral retina. For instance, a minute gnat flying across the peripheral field of vision is instantaneously detected. On the other hand, the same gnat sitting quietly in the peripheral field of vision remains entirely below the threshold of visual detection.

Transmission of Color Signals by the Ganglion Cells. A single ganglion cell may be stimulated by a number of cones or by only a very few. When all three types of cones—the red, blue, and green types—all stimulate the same ganglion cell, the signal transmitted through the ganglion cell is the same for any color of the spectrum. Therefore, this signal plays no role in the detection of the different colors. Instead, it is a "white" signal.

On the other hand, many of the ganglion cells are excited by only one color type of cone but inhibited by a second type. For instance, this frequently occurs for the red and green cones, red causing excitation and green causing inhibition—or vice versa, that is, green causing excitation and red, inhibition. The same type of reciprocal effect also occurs between blue cones on the one hand and a combination of red and green cones on the other hand, giving a reciprocal excitation inhibition relationship between the blue and yellow colors.

The mechanism of this opposing effect of colors is the following: One color-type cone excites the ganglion cell by the direct excitatory route through a depolarizing bipolar cell, while the other color type inhibits the ganglion cell by the indirect inhibitory route through a horizontal cell or a hyperpolarizing bipolar cell.

The importance of these color-contrast mechanisms is that they represent a mechanism by which the retina itself begins to differentiate colors. Thus each color-coatrast type of ganglion cell is excited by one color but inhibited by the "opponent color." Therefore, the process of color analysis begins in the retina and is not entirely a function of the brain.

REFERENCES

Allansmith, M R: The Eye and Immunology St Louis, C V Mosby, 1983

Callender, R, and Honig, B Resonance Raman studies of visual pigments *Annu Rev Biophys Bioeng*, *6*:33, 1977.

Cervetto, L., and Fuortes, M G F. Excitation and interactions in the retina *Annu Rev Biophys Bioeng*, *7* 229, 1978

Chignell, A H Retinal Detachment Surgery New York, Springer-Verlag, 1979

Cunha-Vaz, J G (ed) The Blood Retinal Barriers New York, Plenum Press, 1980

Davson, H The Physiology of the Eye, 3rd Ed New York, Academic Press, 1972

Daw N W Neurophysiology of color vision *Physiol Rev*, *53* 571, 1973

DeValois, R. L., and Jacobs, G H Neural mechanisms of color vision. *In* Darian-Smith, I, (ed) Handbook of Physiology Sec I, Vol III Bethesda, American Physiological Society, 1984, p 525

Dick, G L Studies in Ocular Anatomy and Physiology Kensington, N S W New South Wales University Press, 1976

Dowling, J E, and Dubin, M W The vertebrate retina *In* Darian Smith, I (ed) Handbook of Physiology Sec I, Vol III Bethesda, American Physiological Society, 1984, p 317

Fatt, I Physiology of the Eye An Introduction to the Vegetative Function Boston, Butterworths, 1978

Favreau O E, and Corballis, M C Negative aftereffects in visual perception *Sci Am* 235(6) 42, 1976

Fine, B S, and Yanoff, M Ocular Histology A Text and Atlas. Hagerstown, Md, Harper & Row, 1979

Freeman, R D (ed) Developmental Neurobiology of Vision. New York, Plenum Press, 1979

Friedlaender, M H Allergy and Immunology of the Eye Hagerstown, Md, Harper & Row, 1979

Hillman, P, *et al* Transduction in invertebrate photoreceptors Role of pigment bistability *Physiol Rev*, *63* 668, 1983

Kaneko, A Physiology of the retina. *Annu Rev Neurosci*, *2* 169, 1979

Kanski, J J (ed) BIMR Ophthalmology Vol 1 Disorders of the Vitreous, Retina, and Choroid Woburn, MA, Butterworths, 1983

Kohner, E M (ed) Diabetic Retinopathy Boston, Little, Brown, 1978

Land, E H The retinex theory of color vision *Sci Am*, *237*(6) 108, 1977

MacNichol, E F, Jr Three pigment color vision *Sci Am*, *211* 48, 1964

Marks, W B, *et al* Visual pigments of single primate cones. *Science*, *143* 1181, 1964

Michael, C R. Color vision *N Engl J Med*, *288* 724, 1973

Michaelson, I C Textbook of the Fundus of The Eye. New York, Churchill Livingstone, 1980

Ming, A. L. S, and Constable, I. J Colour Atlas of Ophthalmology Boston, Houghton Mifflin, 1979

Moses, R A. Adler's Physiology of the Eye, Clinical Application 7th Ed St. Louis, C V Mosby, 1981

Padgham, C A., and Saunders, J. E The Perception of Light and Color New York, Academic Press, 1975

Retina and Vitreous Rochester, Minn., American Academy of Ophthalmology, 1978

Rushton, W A H.: Visual pigments and color blindness Sci Am, *232*(3) 64, 1975

Schepens, C L: Retinal Detachment and Allied Diseases Philadelphia, W B Saunders Co, 1983

Sherman, S M., and Spear, P D Organization of visual pathways in normal and visually deprived cats. *Physiol Rev*, *62* 738, 1982

Wolf, G Multiple functions of vitamin A. *Physiol Rev*, *64* 873, 1984

Yannuzzi, L. A., *et al.* (eds) The Macula. A Comprehensive Text and Atlas. Baltimore, Williams & Wilkins, 1978

Young, R. W Proceedings. Bingeness and renewal of visual cell outer segment membranes *Exp Eye Res*, *18* 215, 1974

Zimmerman, T. J, *et al* Advances in ocular pharmacology *Annu Rev. Pharmacol. Toxicol*, *20* 415, 1980

Zinn, K M, and Marmor, M. F (eds.) The Retinal Pigment Epithelium Cambridge, Mass., Harvard University Press, 1979.

60

The Eye: III. Central Neurophysiology of Vision

THE VISUAL PATHWAY

Figure 60–1 illustrates the visual pathway from the two retinas to the *visual cortex* After impulses leave the retinas they pass backward through the *optic nerves*. At the *optic chiasm* all the fibers from the nasal halves of the retinas cross to the opposite side where they join the fibers from the opposite temporal retinaa to form the *optic tracts*. The fibers of each optic tract aynapse in the *laterol geniculate body* and from here, the *geniculocalcarine fibers* paas through the *optic rodiation*, or *geniculocalcarine tract*, to the *primary visual cortex* in the calcarine area of the occipital lobe

In addition, visual fibers also pass to other areas of the brain: (1) from the optic tracts to the *suprochiasmatic nucleus of the hypothalamus*, presumably for controlling circadian rhythms; (2) into the *pretectal nuclei*, for control of fixation of the eyes on objects of importance and also for activa the pupillary light reflex; (3) into the *sup colliculus*, for control of bilateral simultan movement of the two eyes; and (4) into the *pul* as a secondary visual pathway either directly f the optic tracts or indirectly from the supe colliculus.

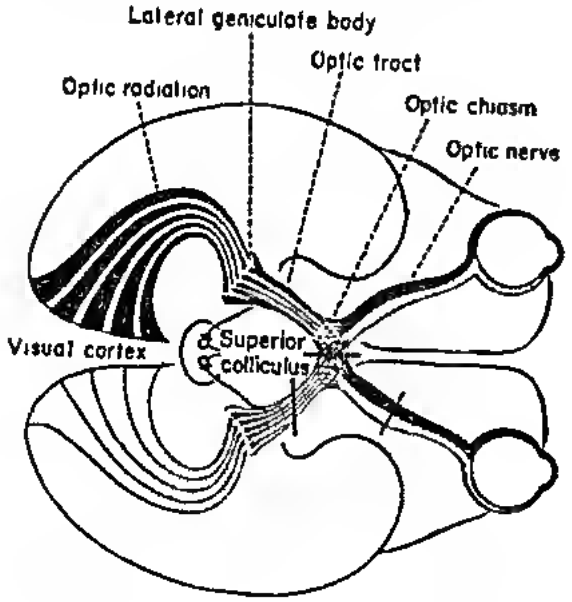


Figure 60–1. The visual pathways from the eyes to the visual cortex. (Modified from Polyak. The Retina. University of Chicago Press.)

THREE DIFFERENT TYPES OF RETINAL GANGLIAL CELLS AND THEIR PROJECTIONS

The ganglion cells of the retina are not all same type either anatomically or functiona They are mainly divided into three separate ty large, medium, and small.

About 55 per cent of tha ganglion cells ar medium size, 10 to 15 microns in diameter. E of these is excited by a small, circumscribed gr of rods or cones or both. This is especially tru the ganglion cells in the foveal region where e one connects with only a few cones. These mediu aized ganglion cells then, after relaying their s nals in the lateral geniculate body, transmit visual pattern to the visual cortex.

The small ganglion cells, less than 10 micr in diameter and representing about 40 per cent all the ganglion cells, have very thin dendri that spread widely in the retina. In turn, th axons project directly to the superior colliculus a adjacent regions of the pretectal nuclei. The small ganglion cells are especially important detecting movement in the visual image, and response they cause turning of the eyes and he toward the moving object.

The very large ganglion cells, some having ameters as large as 35 microns, represent only per cent of all the ganglion cells. They have a ve broad dendritic field in the retina and therefo transmit general information into the visual pat way. They project to both the lateral geniculat body and to the superior colliculus.

Thus, specific features of visual information b gin to be dissected from other features before th

sual signals leave the retina; then the different pes of information are transmitted through different pathways to different brain areas.

unction of the Lateral Geniculate Body

Each lateral geniculate body is located laterally t the dorsal end of the thalamus. In this body the *ptic tract terminates*, and from here the visual ignals are relayed through the *geniculocolcarine ract to the visual cortex.*

Each lateral geniculate body is composed of six uclear layers. Layers 2, 3, and 5 (from ventral to orsal) receive signals from the temporal portion f the ipsilateral retina, whereas layers 1, 4, and receive signals that cross in the optic chiasm rom the nasal retina of the opposite eye.

The respective retinal areas of the two eyes onnect with neurons that are approximately superimposed over each other in the paired layers, nd the layers in turn connect with each other by hort axons. Therefore, with a little imagination, ne can postulate that the adjacent layers of the ateral geniculate body perhaps play roles in two mportant processes: (1) *fusion of vision*, which neans appropriate control of the directions of gaze f the two eyes so that the two retinal images will superimpose exactly onto each other, and (2) *stereoscopic depth perception*, which depends on comparing the visual images of the two eyes and determining their slight differences, as was discussed in Chapter 58.

Relay Function of the Lateral Geniculate Body. The lateral geniculate body does not change the "pattern" of the visual image significantly as it relays this image from the retina to the visual cortex. Also, the neurons in the lateral geniculate body respond in almost the same manner as the medium-sized ganglion cells of the retina. That is, they respond very strongly to contrast borders in the retinal image, to spots of light, and to opponent colors.

It is not known why the lateral geniculate body has three entirely different pairs of neuronal layers. However, the signals in layers 1 and 2 are mainly related to black and white vision, whereas the color signals occur mainly in layers 3 and 4, and layers 5 and 6.

FUNCTION OF THE PRIMARY VISUAL CORTEX

The ability of the visual system to detect spatial organization of the visual scene—that is, to detect the forms of objects, brightness of the individual parts of the objects, shading, and so forth—depends on function of the *primary visual cortex*, the anatomy of which is illustrated in Figure 60–2. This area lies mainly in the calcarine fissure located bilaterally on the medial aspect of each occipital

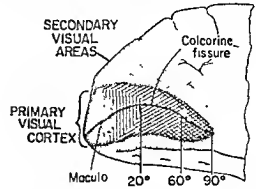


Figure 60–2. The visual cortex.

cortex. Specific points of the retina connect with specific points of the visual cortex, the right halves of the two respective retinas connecting with the right visual cortex and the left halves connecting with the left visual cortex. The macula is represented at the occipital pole of the visual cortex and the peripheral regions of the retina are represented in concentric circles farther and farther forward from the occipital pole. The upper portion of the retina is represented superiorly in the visual cortex, and the lower portion, inferiorly. Note the large area of cortex receiving signals from the macular region of the retina. It is in this region that the fovea is represented, which gives the highest degree of visual acuity. Based on foveal area, the fovea has 35 times as much representation in the primary visual cortex as do the peripheral portions of the retina.

DETECTION OF LINES AND BORDERS BY THE PRIMARY VISUAL CORTEX

If a person looks at a blank wall, only a few neurons of the primary visual cortex will be stimulated, whether the illumination of the wall is bright or weak. Therefore, the question must be asked. What does the visual cortex do? To answer this question, let us now place on the wall a large solid cross such as that illustrated to the left in Figure 60–3. To the right is illustrated the spatial pattern of the greater majority of the excited neurons that one finds in the visual cortex. *Note that*

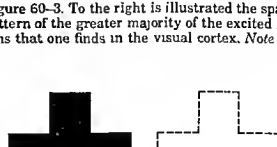


Figure 60–3. Pattern of excitation occurring in the visual cortex in response to a retinal image of a dark cross

the areas of excitation occur along the sharp borders of the visual pattern Thus, by the time the visual signal is recorded in the primary visual cortex, it is concerned mainly with the *contrasts* in the visual scene rather than with the flat areas. At each point in the visual scene where there is a change from dark to light or light to dark, the corresponding area of the primary visual cortex becomes stimulated The intensity of stimulation is determined by the *gradient of contrast*. That is, the greater the sharpness in the contrast border and the greater the difference in intensities between the light and dark areas, the greater the degree of stimulation.

Thus, the *pattern of contrasts* in the visual scene is impressed upon the neurons of the visual cortex, and this pattern has a spatial orientation roughly the same as that of the retinal image

Detection of Orientation of Lines and Borders—The Neuronal Columns

Not only does the visual cortex detect the *existence* of lines and borders in the different areas of the retinal image, but it also detects the *orientation* of each line or border—that is, whether it is a vertical or horizontal line or border, or lies at some degree of inclination The mechanism of this effect is the following:

In each small area of the visual cortex the neurons are arranged in columns, each column containing many neurons and having a diameter of 200 to 500 microns, extending downward from the surface of the cortex through its six layers. The signals arriving in the cortex from the lateral geniculate body terminate in layer 4, and from here secondary signals spread upward or downward in the column

Within each column of neurons, the neurons of layer 4 respond to lines that (a) are located at single positions on the retina and (b) are oriented in a specific direction Thus, these neurons are line detectors that are both direction-specific and position-specific. They are called *simple cells*.

Detection of Line Orientation Even When the Line Is Displaced Laterally. As the signal progresses farther away from layer 4 of each column, some of the neurons now respond to lines still oriented in the same direction but not position-specific. That is, the line can be displaced laterally in either direction over a moderate distance, and still the neuron will be stimulated. Therefore, these cells are called *complex cells*.

Detection of Length of Lines. Many of the neurons in the outer layers of each column of cells are stimulated only by lines or borders *of specific lengths* Thus, this is a still higher order for deciphering information from the visual scene; these neurons are called *hypercomplex cells*.

Changing Orientation of the Lines from One Column of Neurons to the Next. When one moves a recording electrode from one colum cells of the primary visual cortex to an adja column, it is seen that now a different oriente af lines is required to excite the column, usu rotated about 10 degrees from those of the column. And, when going to the next column, lines are rotated still another 10 degrees T the separate columns of cells detect images ented in different directions.

The Neuronal Mechanism for Detec Lines. The mechanism by which the neuron the primary visual cortex detect lines rather t spots of light is believed to be the following: l believed that each linear group of ganglion c in the retina project their signals to a si "simple cell" in layer 4 of the primary vis cortex When one ganglion cell of the grou stimulated but the rest are not, the signal inten is not strong enough to excite the simple c However, when the entire group of ganglion c in the linear sequence are stimulated, then signal strength is great enough to excite the sim cell

Now, if we remember that each excitatory g glion cell of the retina is excited by light bu also excited by darkness falling on the retina the sides of the ganglion cell, then it becomes cl that any contrast border in the visual image v cause double excitation of the linear group ganglion cells—that is, (1) direct excitation cau by the light portion of the contrast border, and excitation caused by lateral inhibition from t laterally placed dark portion of the retina. Th it is this organization of the neuronal connectic from the ganglion cells of the retina that allo the simple cells of the primary visual cortex detect lines.

In a similar manner, the complex cells and t hypercomplex cells are believed to receive convei ing excitatory and inhibitory neuronal signa from the lower order neurons, in this way de phering more and more complex information fro the scene.

ANALYSIS OF COLOR BY THE VISUAL CORTEX

Some of the columns of cells in the visual cort respond specifically to colors rather than simply lines. However, as is true in the retina and also i the lateral geniculate body, stimulation of th neurons in these columns usually requires co trasting opponent colors, especially the red-gree opponent colors and the blue-yellow opponer colors. The deciphering of color information l believed also to become more complex at differen levels of neuronal organization. In fact, full appre ciation of color probably does not occur in th primary visual cortex alone, for it has recentl been found in human beings that lesions on th ventral surface of the occipital lobe can cause lac

color reception. It is suggested, therefore, that s area might be a final region for color signal cessing.

nteraction of Visual Signals from the Two Iferent Eyes in the Primary Visual Cortex. om our earlier description of the lateral genic-te body, it will be recalled that the visual nals from the two separate eyes are relayed ough separate lateral geniculate body neuronal ers. These signals from the two eyes are still arated from each other when they arrive in the mary visual cortex. The cortex is interlaced h zebra-like stripes of neurons, each stripe out 0.5 mm wide; the signals from one eye enter ry other stripe, and the signals from the other the opposite set of stripes. The signals from two eyes remain separated from each other en they excite the layer 4 neurons; this is also e for excitation of almost all the simple cells t detect simple line orientation. However, by time the complex, and especially the hypercom-x, cells are excited, many of the signals from two separate eyes have by then begun to verge. As we shall see later in the chapter, en these signals are not "in register"—that is, both excited or inhibited together—*interference terns* develop and excite still other cells It is sumed that these patterns are the basis for trolling fusion of the visual images from the eyes, and they are likely also the basis of reopsis.

RCEPTION OF LUMINOSITY

Very little is known about the way in which the ain detects tho level of luminosity. However, it believed that this results at least partly from effect that luminosity has of increasing the ensity of the visual contrasts caused by lines, rders, moving objects, and opponent colors in visual scene. That is, the greater the light ensity, the greater are the degrees of contrast d therefore the more strongly the visual cortex stimulated.

In addition, some of the ganglion cells of the lina do respond to luminosity, especially some of e very small ganglion cells. These transmit lu-inosity signals into the pretectal region of the ain stem to control the size of the pupil. It is ssible that this same information is used to help e person appreciate the level of illumination.

FECT OF REMOVING THE PRIMARY SUAL CORTEX

Removal of the primary visual cortex in the man being always causes loss of conscious vi-on. However, psychological studies have demon-rated that such persons can still react subcon-ously to many aspects of vision, especially to anges in light intensity, and to crude aspects of form as well. The reactions include turning the eyes, turning the head, avoidance, and so on. In some low-level mammals a large share of the vision is still preserved even when the primary visual cortex is completely removed. This vision is believed to be subserved by neuronal pathways from the optic tracts to the superior colliculi and thence to the secondary visual cortical areas.

TRANSMISSION OF VISUAL INFORMATION INTO OTHER REGIONS OF THE CEREBRAL CORTEX

The primary visual cortex is located in the *calcarine fissure area*, mainly on the medial aspect of each hemisphere but also spreading slightly over the occipital pole It is known as *Brodmann area 17* of the cortex, and it is also called the *striate area* because of its striated appearance to the naked eye.

Signals from the striate area project laterally in the occipital cortex into area 18 and then into area 19, as illustrated in Figure 60–4. These areas are called the *secondary visual cortex*, the *peristriate area*, or *visual association areas*. In reality, they are simply the loci for additional processing of visual information.

The visual projection images in areas 18 and 19 are organized into columns of cells in the same manner as described earlier for the primary visual cortex. There are also as many as five different topographic representations of the visual field as found in the primary visual cortex. However, the neuronal cells are mainly of the *hypercomplex type* and respond to more complex patterns than do those in the primary visual cortex. Presumably, these progressively more complex interpretations eventually decode the visual information, giving the overall impression of the visual scene that the person is observing

Projection of Visual Informatioo into the Temporal Cortex. From areas 18 and 19, visual signals next proceed to the posterior portion of the temporal lobes, Brodmann areas 20 and 21. In these areas simple visual patterns such as lines, borders, angles, and so forth fail to cause excitation of specific neurons. The degree of integration of visual information at this level, such as interpretation of letters or words, is presumably much higher

Effects of Destruction of the Secondary Visual Cortex. Human beings who have destructive lesions of any of the visual association areas, areas 18 and 19 in the occipital cortex or areas 20 and 21 in the temporal

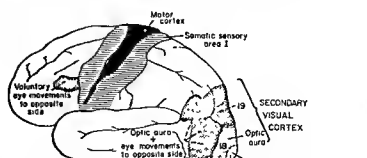


Figure 60–4. The visual association fields and the cortical areas for control of eye movements.

cortex, have difficulty with certain types of visual perception and visual learning. For instance, destruction in areas 18 and 19 generally makes it difficult to perceive form, such as shapes of objects, their sizes, and their meanings It is believed that this type of lesion in the dominant hemisphere can cause the abnormality known as *dyslexia* or *word blindness*, which means that the person has difficulty understanding the meanings of words that are seen

Destruction of the temporal projections of the visual system makes it especially difficult for animals, and presumably for human beings as well, to learn tasks that are based upon visual perceptions For instance, a person might be able to see a plate of food perfectly well but be unable to utilize this visual information to direct a fork toward the food Yet, when the person feels the plate with the hand, the fork can be directed accurately by stereotactic information from the somesthetic cortex.

THE FIELDS OF VISION; PERIMETRY

The *field of vision* is the area seen by an eye at a given instant. The area seen to the nasal side is called the *nasal field of vision*, and the area seen to the lateral side is called the *temporal field of vision*

To diagnose blindness in specific portions of the retinas, one charts the field of vision for each eye by a process known as *perimetry* This is done by having the subject look with one eye toward a central spot directly in front of the eye Then a small dot of light or a small object is moved back and forth in all areas of the field of vision, both laterally and nasally and upward and downward, and the person indicates when the spot of light or object can be seen and when it cannot At the same time, a chart (Fig. 60–5) is made for the eye, showing the areas in which the subject can and cannot see the spot Thus, the field of vision is plotted

In all perimetry charts, a *blind spot* caused by lack of rods and cones in the retina over the *optic disc* is found approximately 15 degrees lateral to the central point of vision, as illustrated in the figure

Abnormalities in the Fields of Vision. Occasionally blind spots are found in other portions of the field of vision besides the optic disc area. Such blind spots are called *scotomata;* they frequently result from allergic reactions in the retina or from toxic conditions, such as lead poisoning or even excessive use of tobacco

Still another condition that can be diagnosed by perimetry is *retinitis pigmentosa*. In this disease, portions of the retina degenerate and excessive melanin pigment deposits in the degenerated areas Retinitis pigmentosa generally causes blindness in the peripheral field of vision first and then gradually encroaches on the central areas

Effect of Lesions in the Optic Pathway on Fields of Vision. One of the most important uses of perimetry is for localization of lesions in the visual nervous pathway Lesions in the optic nerve, in the optic chiasm, in the optic tract, and in the geniculocalcarine tract all cause blind areas in the visual fields, and the "patterns" of these blind areas often indicate the location of the lesion

Destruction of an entire *optic nerve* obviously causes blindness of the respective eye. Destruction of the *optic chiasm*, as shown by the longitudinal line across the chiasm in Figure 60–1, prevents the passage of impulses from the nasal halves of the two retinae to the opposite optic tracts Therefore, the nasal halves are both blind, which means that the person is blind in both temporal fields of vision *because the image of the field of vision is inverted on the retina*, this condition is called *bitemporal hemianopsia* Such lesions frequently result from tumors of the adenohypophysis pressing upward on the optic chiasm

Interruption of an *optic tract*, which is also shown by a line in Figure 60–1, denervates the corresponding half of each retina on the same side as the lesion, and as a result, neither eye can see objects to the opposite side This condition is known as *homonymous hemianopsia* Destruction of the *optic radiation* or the visual cortex on one side also causes homonymous hemianopsia A common condition that destroys the visual cortex is thrombosis of the posterior cerebral artery, which infarcts the occipital cortex except for the foveal area, thus sparing central vision.

One can differentiate a lesion in the optic tract from a lesion in the geniculocalcarine tract or visual cortex by determining whether impulses can still be transmitted into the pretectal nuclei to initiate a pupillary light reflex

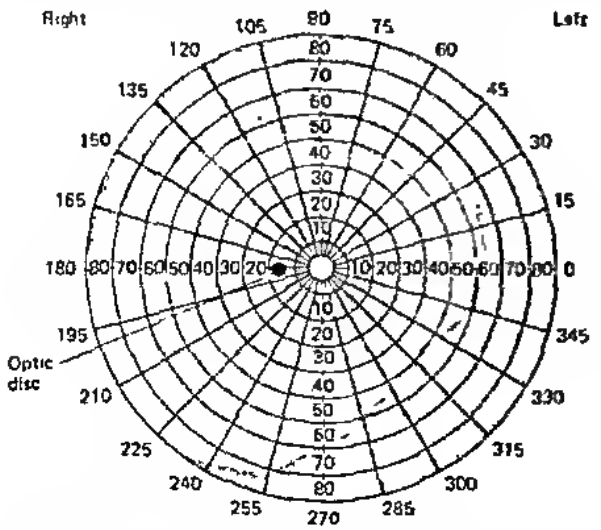


Figure 60–5. A perimetry chart, showing the field of vision for the left eye

EYE MOVEMENTS AND THEIR CONTROL

To make use of the abilities of the eye, almost equally as important as the system for interpretation of the visual signals from the eyes is the cerebral control system for directing the eyes toward the object to be viewed.

Muscular Control of Eye Movements. The eye movements are controlled by three separate pairs of muscles, shown in Figure 60–6: (1) the *medial* and *lateral recti*, (2) the *superior* and *inferior recti*, and (3) the *superior* and *inferior oblique*. The medial and lateral recti contract reciprocally to move the eyes from side to side. The superior and inferior recti contract reciprocally to move

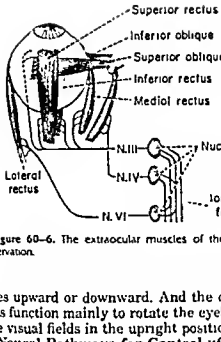


Figure 60–6. The extraocular muscles of the eye and their nervation.

ies upward or downward. And the oblique mus-es function mainly to rotate the eyeballs to keep ie visual fields in the upright position

Neural Pathways for Control uf Eye Move-ients. Figure 60–6 also illustrates the nuclei of ie third, fourth, and sixth cranial nerves and ieir innervation of the ocular muscles Shown, ıo, are the interconnections among these three uclei through the *medial longitudinal fasciculus* ıther by way of this fasciculus or by way of other losely associated pathways, each of the three sets f muscles to each eye is *reciprocally* innervated so that one muscle of the pair relaxes while the other contracts.

Figure 60–7 illustrates cortical control of the oculomotor apparatus, showing spread of signals from the occipital visual areas through occipitotectal and occipitocollicular tracts into the pretectal and superior colliculus areas of the brain stem. In addition, a frontotectal tract passes from the frontal cortex into the pretectal area. From both the pretectal and the superior colliculus areas, the oculomotor control signals then pass to the nuclei of the oculomotor nerves. Strong signals are also transmitted into the oculomotor system from the vestibular nuclei by way of the medial longitudinal fasciculus

FIXATION MOVEMENTS OF THE EYES

Perhaps the most important movements of the eyes are those that cause the eyes to "fix" on a discrete portion of the field of vision.

Fixation movements are controlled by two different neuronal mechanisms The first of these allows the person to move his eyes voluntarily to find the object upon which he wishes to fix his vision; this is called the *voluntary fixation mechanism* The second is an involuntary mechanism that holds the eyes firmly on the object once it has been found; this is called the *involuntary fixation mechanism.*

The voluntary fixation movements are controlled by a small cortical field located bilaterally in the premotor cortical regions of the frontal lobes, as illustrated in

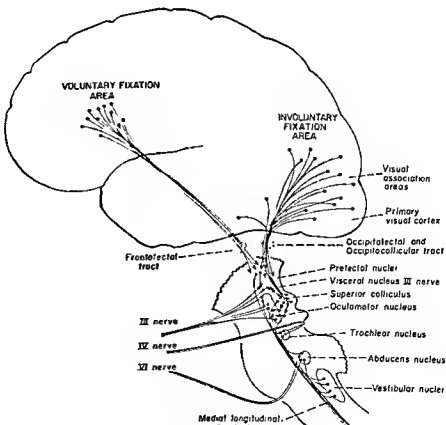


Figure 60–7. Neural pathways for control of conjugate movement of the eyes.

Figure 60–7. Bilateral dysfunction or destruction of these areas makes it difficult or almost impossible for the person to "unlock" the eyes from one point of fixation and then move them to another point. It is usually necessary for the person to blink eyes or put a hand over the eyes for a short time, which then allows the eyes to be moved

On the other hand, the fixation mechanism that causes the eyes to "lock" on the object of attention once it is found is controlled by the *secondary visual area of the occipital cortex*—mainly area 19—which is also illustrated in Figure 60–7. When this area is destroyed bilaterally, an animal has difficulty keeping its eyes directed toward a given fixation point or becomes completely unable to do so

To summarize, the posterior eye fields automatically "lock" the eyes on a given spot of the visual field and thereby prevent movement of the image across the retina To unlock this visual fixation, voluntary impulses must be transmitted from the "voluntary" eye fields located in the frontal areas.

Mechanism of Fixation. Visual fixation results from a negative feedback mechanism that prevents the object of attention from leaving the foveal portion of the retina. The eyes even normally have three types of continuous but almost imperceptible movements (1) a *continuous tremor* at a rate of 30 to 80 cycles per second caused by successive contractions of the motor units in the ocular muscles, (2) a *slow drift* of the eyeballs in one direction or another, and (3) sudden *flicking movements* which are controlled by the involuntary fixation mechanism When a spot of light has become fixed on the foveal region of the retina, the tremorous movements cause the spot to move back and forth at a rapid rate across the cones, and the drifting movements cause it to drift slowly across the cones However, each time the spot of light drifts as far as the edge of the fovea, a sudden reflex reaction occurs, producing a flicking movement that moves the spot away from this edge back toward the center Thus, whenever the image drifts away from the point of fixation, an automatic response moves the image back toward the central portion of the fovea These drifting and flicking motions are illustrated in Figure 60–8, which shows by the dashed lines the slow drifting across the retina and by the solid lines the flicks that keep the image from leaving the foveal region

Fixation of the Eyes on Important Visual Highlights—Role of the Superior Colliculi. The eyes have an automatic capability for instantaneously *fixing* on an important highlight in the visual field However, this capability is mostly lost when the superior colliculi are destroyed. The signals for fixation originate in the visual areas of the occipital cortex, primarily in the secondary visual areas, then pass to the superior colliculi, probably from there to reticular areas around the oculomotor nuclei, and thence into the motor nuclei themselves

Function of the Superior Colliculi in Turning of the Eyes or the Head Toward a Visual Disturbance. The superior colliculi are also important in causing sudden turning of the eyes to the side when a flash of light or some other sudden visual disturbance occurs on that side. In addition, signals are transmitted from the superior colliculi through the medial longitudinal fasciculus to other areas of the brain stem to cause turning of the whole head and perhaps even of the whole body toward the direction of the light Other types of disturbances besides visual disturbances, such as strong sounds or even stroking the side of the body, will cause

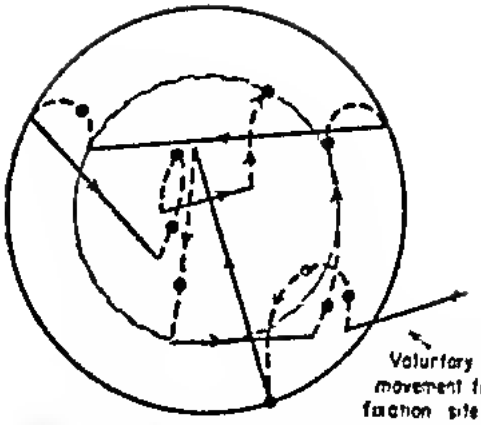


Figure 60–8. Movements of a spot of light on the fovea showing sudden "flicking" movements to move the spot back toward the center of the fovea whenever it drifts to the fovea edge (The dashed lines represent slow drifting movements and the solid lines represent sudden flicking movements) (Modified from Whitteridge: Handbook of Physiology, Vol. 2, Sec. 1 Baltimore Williams & Wilkins, 1960, p 1089)

similar turning of the eyes, head, and body if the superior colliculi are intact. This effect, also, is absent or severely disturbed when the superior colliculi are destroyed Therefore, it is frequently said that the superior colliculi play an important role in orienting the eyes, the head, and the body with respect to external signals—visual signals, auditory signals, somatic signals, and perhaps even others.

Saccadic Movement of the Eyes. When the visual scene is moving continually before the eyes, such as when a person is riding in a car or turning around, the eyes fix on one highlight after another in the visual field, jumping from one to the next at a rate of two to three jumps per second The jumps are called *saccades*, and the movements are called *opticokinetic movements*. The saccades occur so rapidly that not more than 10 per cent of the total time is spent in moving the eyes, 90 per cent of the time being allocated to the fixation sites. Also, the brain suppresses the visual image during the saccades so that one is completely unconscious of the movements from point to point.

Saccadic Movements During Reading. During the process of reading, a person usually makes several saccadic movements of the eyes for each line. In this case the visual scene is not moving past the eyes, but the eyes are trained to scan across the visual scene to extract the important information Similar saccades occur when a person observes a painting, except that the saccades occur in one direction after another from one highlight of the painting to another, then another, and so forth

Fixation on Moving Objects—"Pursuit Movements." The eyes can also remain fixed on a moving object, which is called *pursuit movement*. A highly developed cortical mechanism automatically detects the course of movement of an object and then gradually develops a similar course of movement of the eyes. For instance, if an object is moving up and down in a wavelike form at a rate of several times per second, the eyes at first may be completely unable to fixate on it. However, after a second or so the eyes begin to jump coarsely in approximately the same pattern of movement as that of the object. Then after a few more seconds, the eyes develop progressively smoother and smoother

movements and finally follow the course of movement almost exactly. This represents a high degree of automatic, subconscious computational ability by the cerebral cortex.

Vestibular Control of Eye Movements. Another type of eye movement is elicited by stimulation of the vestibular apparatus. The vestibular nuclei are connected directly with the brainstem nuclei that control ocular movements, and, whenever the head is accelerated in a vertical, longitudinal, lateral or angular direction, an immediate compensatory motion of the eyes occurs in the opposite direction. This allows the eyes to remain fixed on an object of attention despite rapid movements of the body or head.

Vestibular control of the eyes is especially valuable when a person is subjected to jerky motions of the body. For instance, when a person with bilateral destruction of the vestibular apparatuses rides over rough roads, fixing the eyes on the road or on any horizontal scene is extremely difficult. The optickinetic type of movement is not capable of keeping the eyes fixed under such conditions because it has a latent period of about one-fifth second before the direction of movement of the visual scene can be detected and followed by the eyes. On the other hand, the vestibular type of eye movement has a short latent period, measured in thousandths of a second rather than one-fifth second.

Pathologic Nystagmus. Occasionally, abnormalities occur in the control system for eye movements that cause continuous nystagmus (back and forth movements) despite the fact that neither the visual scene nor the body is moving. This is likely to occur when one of the vestibular apparatuses is damaged or when severe damage is sustained in the deep nuclei of the cerebellum. This was discussed further in Chapter 53.

Another pathologic type of eye movement that is sometimes called nystagmus occurs when the foveal regions of the two eyes have been destroyed or when the vision in these areas is greatly weakened. In such a condition the eyes attempt to fix the object of attention on the foveae but always overshoot the mark because of foveal insensitivity. Therefore, they oscillate back and forth but never achieve foveal fixation.

FUSION OF THE VISUAL IMAGES

To make the visual perceptions more meaningful and also to aid in depth perception by the mechanism of stereopsis, which was discussed in Chapter 58, the visual images in the two eyes normally *fuse* with each other on "corresponding points" of the two retinas.

The visual cortex plays a very important role in fusion. It was pointed out earlier in the chapter that corresponding points of the two retinas transmit visual signals to different neuronal layers of the lateral geniculate body, and these signals in turn are relayed to parallel stripes of neurons in the visual cortex. Interactions occur between the stripes of cortical neurons; these cause *interference patterns of excitation* in some of the local neuronal cells when the two visual images are not precisely "in register"—that is, not precisely fused. This excitation presumably provides the signal that is transmitted to the oculomotor apparatus to cause convergence or divergence or rotation of the eyes so that fusion can be re-established. Once the corresponding points of the retinas are precisely in register with each other, the excitation of the specific cells in the visual cortex disappears.

The Neural Mechanism for Stereopsis. The visual images that appear on the retina during the process of stereopsis were discussed in Chapter 58. It was pointed out that because the two eyes are a little more than 2 inches apart the images on the two retinas are not exactly the same. The closer the object is to the eye the greater is the disparity between the two images. Consequently, it is impossible for all corresponding points in the visual image to be in complete register at the same time. Furthermore, the nearer the object is to the eye, the less the degree of register. Here again, specific cells in the primary visual cortex are believed to become excited in the areas of the visual field where highlights are out of register. Presumably, this excitation is the source of the signals for detection of the distance of the object in front of the eyes, this mechanism is called *stereopsis*.

Strabismus. Strabismus, which is also called *squint* or *cross-eyedness*, means lack of fusion of the eyes in one or more of the coordinates described above. Three basic types of strabismus are illustrated in Figure 60-9: *horizontal strabismus, vertical strabismus,* and *torsional strabismus.* However, combinations of two or even of all three of the different types of strabismus often occur.

Strabismus is perpetuated by an abnormal "set" of the fusion mechanism of the visual system. That is, in the early efforts of the child to fixate the two eyes on the same object, one of the eyes fixates satisfactorily while the other fails to fixate, or they both fixate satisfactorily but never simultaneously. Soon, the patterns of conjugate movements of the eyes become abnormally "set" so that the eyes never fuse.

Frequently, some abnormality of the eyes contributes to the failure of the two eyes to fixate on the same point. For instance, if at birth one eye has poor vision in comparison with the other, the good eye tends to fixate on the object of attention while the poor eye might never do so. Also, in hypermetropic infants, intense impulses must be transmitted to the ciliary muscles to focus the eyes, and some of these impulses overflow into the oculomotor nuclei to cause simultaneous convergence of the eyes, as will be discussed later. As a result, the child's fusion mechanism becomes "set" for continual inward deviation of the eyes.

Suppression of Visual Image from a Repressed Eye. In many patients with strabismus the eyes alternate in fixing on the object of attention. However, in other patients, one eye alone is used all the time while the other eye becomes repressed and is never used for vision. The vision in the repressed eye develops only slightly, usually remaining 20/400 or less. If the dominant eye then becomes blinded, vision in the repressed eye can develop only to a slight extent in the adult but far more in young children. This illustrates that visual acuity is highly dependent on proper development of the central synaptic connections from the eyes.

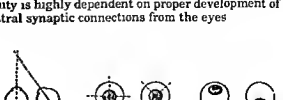


Figure 60-9. The three basic types of strabismus.

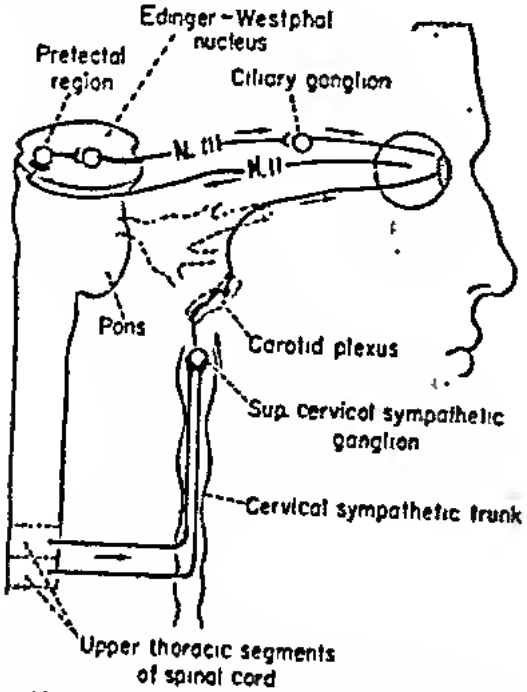


Figure 60–10. Autonomic innervation of the eye showing also the reflex arc of the light reflex. (Modified from Ranson and Clark Anatomy of the Nervous System Philadelphia, W B Saunders Company 1959)

AUTONOMIC CONTROL OF ACCOMMODATION AND PUPILLARY APERTURE

The Autonomic Nerves to the Eyes. The eye is innervated by both parasympathetic and sympathetic fibers, as illustrated in Figure 60–10. The parasympathetic preganglionic fibers arise in the *Edinger-Westphal nucleus* (the visceral nucleus of the third nerve) and then pass in the *third nerve* to the *ciliary ganglion*, which lies about 1 centimeter behind the eye. Here the preganglionic fibers synapse with postganglionic parasympathetic neurons that, in turn, send fibers through the *ciliary nerves* into the eyeball. These nerves excite the ciliary muscle and the sphincter of the iris

The sympathetic innervation of the eye originates in the *intermediolateral horn cells* of the first thoracic segment of the spinal cord From here, sympathetic fibers enter the sympathetic chain and pass upward to the *superior cervical ganglion* where they synapse with postganglionic neurons. Fibers from these spread along the carotid artery and successively smaller arteries until they reach the eye. There the sympathetic fibers innervate the radial fibers of the iris as well as several extraocular structures around the eye, which are discussed shortly in relation to Horner's syndrome. Also, they supply very weak innervation to the ciliary muscle.

CONTROL OF ACCOMMODATION (FOCUSING THE EYES)

The accommodation mechanism—that is, the ism that focuses the lens system of the eye—is essential for a high degree of visual a Accommodation results from contraction or r ation of the ciliary muscle, contraction ca increased strength of the lens system, as expl in Chapter 58, and relaxation causing decre strength. The question that must be answered is How does a person adjust accommodatio keep the eyes in focus all the time?

Accommodation of the lens is regulated negative feedback mechanism that automati adjusts the focal power of the lens for the hig degree of visual acuity. When the eyes have fixed on some far object and then suddenly fi a near object, the lens accommodates for acuity of vision usually within one second. Ti the precise control mechanism that causes rapid and accurate focusing of the eye is unclear, some of the known features of the m anism are the following:

First, when the eyes suddenly change the tance of the fixation point, the lens always chan its strength in the proper direction to achie new state of focus In other words, the lens d *not hunt* back and forth on the two sides of fo in an attempt to find the focus.

Second, different types of clues that can help lens change its strength in the proper direct include the following (1) *Chromatic aberrat* appears to be important. That is, the red light r focus slightly posteriorly to the blue light ra The eyes appear to be able to detect which of th two types of rays is in better focus, and this el relays information to the accommodation mech nism whether to make the lens stronger or weak (2) When the eyes fixate on a near object they al converge toward each other. The neural mech nisms for *convergence cause a simultaneous sign to strengthen the lens of the eye.* (3) *Since the fov is a depressed area, the clarity of focus in the dep of the fovea versus the clarity of focus on the edg will be different* It has been suggested that th also gives clues as to which way the strength the lens needs to be changed. (4) It has been foun that *the degree of accommodation of the lens osc lates slightly* all of the time, at a frequency up t two times per second It has been suggested tha the visual image becomes clearer when the oscil lation of the lens strength is in the appropriat direction and poorer when the lens strength is i the wrong direction. This could give a rapid cu as to which way the strength of the lens needs t change to provide appropriate focus.

It is presumed that the cortical areas that contro accommodation closely parallel those that contro fixation movements of the eyes, with final integration of the visual signals in areas 18 and 19 and transmission of motor signals to the ciliary muscle through the pretectal area and Edinger-Westphal nucleus.

CONTROL OF THE PUPILLARY DIAMETER

Stimulation of the parasympathetic nerves excites the pupillary sphincter, thereby decreasing

the pupillary aperture; this is called *miosis*. On the other hand, stimulation of the sympathetic nerves excites the radial fibers of the iris and causes pupillary dilatation, which is called *mydriasis*.

The Pupillary Light Reflex. When light is shone into the eyes the pupils constrict, a reaction that is called the *pupillary light reflex*. The neuronal pathway for this reflex is illustrated in Figure 60–10. When light impinges on the retina, the resulting impulses pass through the optic nerves and optic tracts to the pretectal nuclei. From here, impulses pass to the *Edinger-Westphal nucleus* and finally back through the *parasympathetic nerves* to constrict the sphincter of the iris. In darkness, the reflex becomes inhibited, which results in dilatation of the pupil.

The function of the light reflex is to help the eye adapt extremely rapidly to changing light conditions, the importance of which was explained in relation to retinal adaptation in the previous chapter. The limits of pupillary diameter are about 1.5 mm on the small side and 8 mm on the large side. Therefore, the range of light adaptation that can be effected by the pupillary reflex is about 30 to 1.

Pupillary Reflexes in Central Nervous System Disease. Certain central nervous system diseases block the transmission of visual signals from the retinas to the Edinger-Westphal nucleus. Such blocks frequently occur as a result of *central nervous system syphilis, alcoholism, encephalitis*, or so forth. The block usually occurs in the pretectal region of the brain stem, though it can also result from destruction of the small afferent fibers in the optic nerves.

The nerve fibers to the Edinger-Westphal nucleus from the pretectal area are of the inhibitory type. Therefore, when they are blocked, the nucleus becomes chronically active, causing the pupils thereafter to remain partially constricted in addition to their failure to respond to light.

When the light reflex is lost, the pupils can still constrict an additional amount if the Edinger-Westphal nucleus is stimulated through some other pathway. For instance, when the eyes fixate on a near object, the signals that cause accommodation of the lens (and also those that cause convergence of the two eyes) cause a mild degree of pupillary constriction at the same time. This is called the *accommodation reflex*. Such a pupil that fails to respond to light but does respond to accommodation and also is very small (an *Argyll Robertson pupil*) is an important diagnostic sign of central nervous system disease—very often syphilis.

Horner's Syndrome. The sympathetic nerves to the eye are occasionally interrupted, and this interruption frequently occurs in the cervical sympathetic chain. This results in *Horner's syndrome*, which consists of the following effects: First, because of interruption of fibers to the pupillary dilator muscle, the pupil remains persistently constricted to a smaller diameter than that of the pupil of the opposite eye. Second, the superior eyelid droops because this eyelid is normally maintained in an open position during the waking hours partly by contraction of a smooth muscle embedded in the lid and innervated by the sympathetics. Therefore, destruction of the sympathetics makes it impossible to open the superior eyelid nearly as widely as normally. Third, the blood vessels on the corresponding side of the face and head become persistently dilated. And, fourth, sweating cannot occur on the side of the face and head affected by Horner's syndrome.

REFERENCES

Anderson, D. R. Testing the Field of Vision. St. Louis, C. V. Mosby, 1963.

Bishop, P. O. Processing of visual information within the retinostriate system. *In* Darnan Smith, I. (ed.) Handbook of Physiology. Sec. 1, Vol. III. Bethesda, American Physiological Society, 1984, p. 341.

Bizzi, E. Eye-head coordination. *In* Brooks, V. B. (ed.). Handbook of Physiology. Sec. 1. Vol. II. Bethesda, American Physiological Society, 1981, p. 1321.

Bittner, E. J. (ed.). Neuroanatomy of the Oculomotor System. New York, Elsevier Science Publishing Co., 1984.

DeValois, R. L. and Jacobs, G. H. Neural mechanisms of color vision. *In* Darnan-Smith, I. (ed.) Handbook of Physiology. Sec. 1. Vol. III. Bethesda, American Physiological Society, 1984, p. 555.

Fraser, S. E. and Hunt, R. K. Retinotectal specificity: Models and experiments in search of a mapping function. *Annu. Rev. Neurosci.*, 3:319, 1980.

Fregnac, Y. and Imbert, M. Development of neuronal selectivity in primary visual cortex of cat. *Physiol. Rev.* 64:325, 1984.

Gilbert, C. D. Microcircuitry of the visual cortex. *Annu. Rev. Neurosci.*, 6:217, 1983.

Hubbell, W. L., and Bownds, M. D. Visual transduction in vertebrate photoreceptors. *Annu. Rev. Neurosci.*, 2:17, 1979.

Hubel, D. H. and Wiesel, T. N. Receptive fields of cells in striate cortex of very young, visually unexperienced kittens. *J. Neurophysiol.*, 26:994, 1963.

Hubel, D. H. and Wiesel, T. N. Cortical and callosal connections concerned with vertical meridian of visual fields in the cat. *J. Neurophysiol.* 30:1561, 1967.

Hubel, D. H. and Wiesel, T. N. Brain mechanisms of vision. *Sci. Am.*, 241:150, 1979.

Mitchell, D. E. and Timney, B. Postnatal development of function in the mammalian visual system. *In* Darnan-Smith, I. (ed.) Handbook of Physiology. Sec. 1. Vol. III. Bethesda, American Physiological Society, 1984, p. 507.

Moses, R. A. Adler's Physiology of the Eye. Clinical Application. 7th Ed. St. Louis, C. V. Mosby, 1981.

Poggio, G. F., and Poggio, T. The analysis of stereopsis. *Annu. Rev. Neurosci.*, 7:379, 1984.

Robinson, D. A. Control of eye movements. *In* Brooks, V. B. (ed.) Handbook of Physiology. Sec. 1, Vol. II. Bethesda, American Physiological Society, 1981, p. 1275.

Schiller, P. H. The superior colliculus and visual function. *In* Darnan-Smith, I. (ed.) Handbook of Physiology. Sec. 1. Vol. III. Bethesda, American Physiological Society, 1984, p. 457.

Schor, C. M. (ed.) Vergence Eye Movements: Basic and Clinical Aspects. Woburn, MA, Butterworth, 1982.

Sherman, S. M., and Spear, P. D. Organization of visual pathways in normal and visually deprived cats. *Physiol. Rev.*, 62:738, 1982.

Simpson, J. I. The accessory optic system. *Annu. Rev. Neurosci.*, 7:13, 1984.

Song, P.-S. Protonion and related photoreceptors. Molecular aspects. *Annu. Rev. Biophys. Bioeng.*, 12:35, 1983.

Sterling, P. Microcircuitry of the cat retina. *Annu. Rev. Neurosci.*, 6:149, 1983.

Walsh, T. J.: Neuro-Ophthalmology: Clinical Signs and Symptoms. Philadelphia, Lea & Febiger, 1978.

Woolsey, C. N. (ed.). Cortical Sensory Organization. Multiple Visual Areas. Clifton, NJ, Humana Press, 1981.

Wurtz, R. H., and Albano, J. E. Visual motor function of the primate superior colliculus. *Annu. Rev. Neurosci.*, 3:189, 1980.

per second, which allows almost full utilization of the energy in the incoming sound waves.

In the absence of the ossicular system and tympanum, sound waves can travel directly through the air of the middle ear and can enter the cochlea at the oval window. However, the sensitivity for hearing is then 15 to 20 decibels less than for ossicular transmission—equivalent to a decrease from a loud voice to a very low voice level.

Attenuation of Sound by Contraction of the Stapedius and Tensor Tympani Muscles. When loud sounds are transmitted through the ossicular system into the central nervous system, a reflex occurs after a latent period of 40 to 80 milliseconds to cause contraction of the *stapedius* and *tensor tympani muscles.* The tensor tympani muscle pulls the handle of the malleus inward while the stapedius muscle pulls the stapes outward. These two forces oppose each other and thereby cause the entire ossicular system to develop a high degree of rigidity, thus greatly reducing the ossicular conduction of low frequency sound, mainly frequencies below 1000 cycles per second.

This *attenuation reflex* can reduce the intensity of sound transmission by as much as 30 to 40 decibels, which is about the same difference as that between a loud voice and the sound of a whisper. The function of this mechanism is probably twofold:

1. To *protect* the cochlea from damaging vibrations caused by excessively loud sound. It is mainly low frequency sounds (the ones that are attenuated) that are frequently loud enough to damage the basilar membrane of the cochlea. Unfortunately, because of the 40 or more millisecond latency for reaction of the reflex, the sudden, loud, thunderous sounds that result from explosions can still cause extensive cochlear damage.

2. To *mask* low frequency sounds in loud environments. This usually removes a major share of the background noise and allows a person to concentrate on sounds above 1000 cycles per second frequency. It is in this upper frequency range that most of the pertinent information in voice communication is transmitted.

Another function of the tensor tympani and stapedius muscles is to decrease a person's hearing sensitivity to his or her own speech. This effect is activated by collateral signals transmitted to these muscles at the same time that the brain activates the voice mechanism.

TRANSMISSION OF SOUND THROUGH BONE

Because the inner ear, the *cochlea,* is embedded in a bony cavity in the temporal bone called the bony labyrinth, vibrations of the entire skull can cause fluid vibrations in the cochlea itself. Therefore, under appropriate conditions, a tuning fork or an electronic vibrator placed on any bony protuberance of the skull causes the person to hear the sound. Unfortunately, the energy available even in very loud sound in the air is not sufficient to cause hearing through the bone except when a special electromechanical sound-transmitting device is applied directly to the bone, usually to the mastoid process.

THE COCHLEA

FUNCTIONAL ANATOMY OF THE COCHLEA

The cochlea is a system of coiled tubes, shown in Figure 61–1 and in cross-section in Figures 61–2 and 61–3. It is comprised of three different tubes coiled side by side: the *scala vestibuli,* the *scala media,* and the *scala tympani.* The scala vestibuli and scala media are separated from each other by *Reissner's membrane* (also called the *vestibular membrane),* and the scala tympani and scala media are separated from each other by the *basilar membrane.* On the surface of the basilar membrane lies a structure, the *organ of Corti,* which contains a series of mechanically sensitive cells, the *hair cells.* These are the receptive end-organs that generate nerve impulses in response to sound vibrations.

Figure 61–4 diagrams the functional parts of the uncoiled cochlea for conduction of sound vibrations. First, note that Reissner's membrane is missing from this figure. This membrane is so thin and so easily moved that it does not obstruct the passage of sound vibrations from the scala vestibuli into the scala media at all. Therefore, so far as the conduction of sound is concerned, the scala vestibuli and scala media are considered to be a single chamber. The importance of Reissner's membrane is to maintain a special fluid in the scala media that is required for normal function of the sound-receptive hair cells, as discussed later in the chapter.

Sound vibrations enter the scala vestibuli from

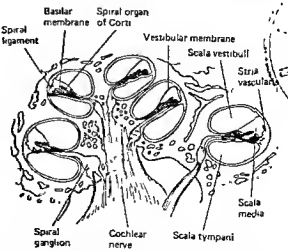


Figure 61–2. The cochlea. (From Goss, C. M. [Anatomy of the Human Body. Lea & Febiger)

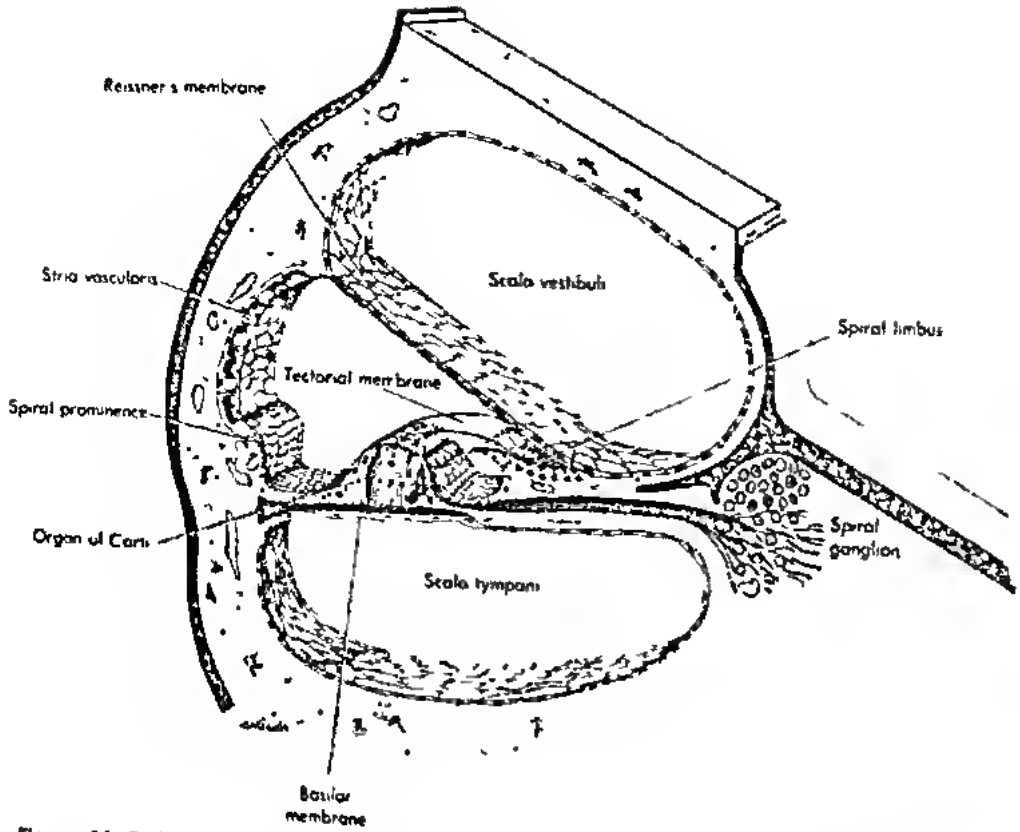


Figure 61–3. A section through one of the turns of the cochlea (Drawn by Sylvia Colard Keene from Bloom and Fawcett A Textbook of Histology Philadelphia, W B Saunders Company 1975)

the faceplate of the stapes at the oval window. The faceplate covers this window and is connected with the window's edges by a relatively loose annular ligament so that it can move inward and outward with the sound vibrations Inward movement causes the fluid to move into the scala vestibuli and scala media, which immediately increases the pressure in the entire cochlea

Note in Figure 61–4 that the distal ends of the scala vestibuli and scala tympani open into each other through an opening at the apex of the cochlea called the *helicotrema* If the stapes moves inward *very slowly*, fluid from the scala vestibuli is pushed through the helicotrema into the scala tympani, and the fluid in the scala tympani then moves backward toward the base of the cochlea where it causes the membrane of the round window to bulge outward. However, if the stapes vibrates inwar and outward rapidly, the fluid simply does nc have time to pass all the way through the helicc trema to the round window, and then back agai to the oval window between each two successiv vibrations Instead, the fluid wave takes a shortcu through the basilar membrane, causing it to bulg back and forth between the scala vestibuli and th scala tympani with each sound vibration. We shal see later that each frequency of sound causes a different "pattern" of vibration in the basilar membrane and that this is the important means by which the sound frequencies are discriminated from each other

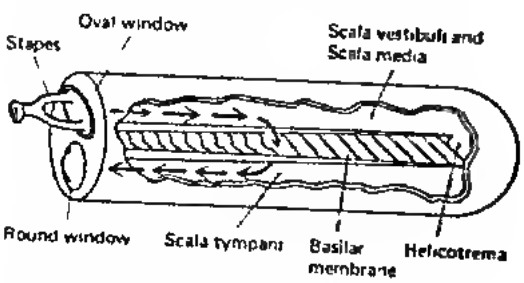


Figure 61–4. Movement of fluid in the cochlea following ward thrust of the stapes.

The Basilar Membrane and Resonance in the Cochlea. The basilar membrane is a fibrous membrane that separates the scala media and the scala tympani. It contains 20,000 to 30,000 *basilar fibers* that project from the bony center of the cochlea, the *modiolus*, toward the outer wall. These fibers are stiff, elastic, reedlike structures that are fixed at their basal ends but not at their distal ends except that the distal ends are embedded in the loose basilar membrane Because the fibers are stiff and also free at one end, they can vibrate like reeds of a harmonica.

The lengths of the basilar fibers increase progressively as one goes from the base of the cochlea to its apex, from a length of approximately 0.04 mm near the oval and round windows to 0 5 mm at the helicotrema, a 12-fold increase in length.

The diameters of the fibers, on the other hand, decrease from the base to the helicotrema, so that their overall stiffness decreases more than 100-fold. As a result, the stiff, short fibers near the base of the cochlea have a tendency to vibrate at a high frequency, whereas the long, limber fibers near the helicotrema have a tendency to vibrate at a low frequency.

In addition to the differences in stiffness of the basilar fibers, they are also differently "loaded" by the fluid mass of the cochlea. That is, when a fiber vibrates back and forth, all the fluid between the vibrating fiber and the oval and round windows must also move back and forth at the same time. For a fiber vibrating near the base of the cochlea (near the two windows), the total mass of moving fluid is slight in comparison with that for a fiber vibrating near the helicotrema. This difference, too, favors high frequency vibration near the windows and low frequency vibration near the tip of the cochlea.

Thus, high frequency resonance of the basilar membrane occurs near the base, and low frequency resonance occurs near the apex because of (1) difference in stiffness of the fibers and (2) difference in "loading."

TRANSMISSION OF SOUND WAVES IN THE COCHLEA—THE "TRAVELING WAVE"

If the foot of the stapes moves inward instantaneously, the round window must also bulge outward instantaneously because the cochlea is bounded on all sides by bony walls. Since the fluid will not have time to move all the way from the oval window through the helicotrema to the round window, the initial effect is to cause the basilar membrane at the very base of the cochlea to bulge in the direction of the round window. However, the elastic tension that is built up in the basilar fibers as they bend toward the round window initiates a wave that "travels" along the basilar membrane toward the helicotrema, as illustrated in Figure 61–5. Figure 61–5A shows movement of a high frequency wave down the basilar membrane; Figure 61–5B, a medium frequency wave; and Figure 61–5C, a very low frequency wave. Movement of the wave along the basilar membrane is comparable to the movement of a pressure wave along the arterial walls, which was discussed in Chapter 19, or it is also comparable to the wave that travels along the surface of a pond.

Pattern of Vibration of the Basilar Membrane for Different Sound Frequencies. Note in Figure 61–5 the different patterns of transmission for sound waves of different frequencies. Each wave is relatively weak at the outset but becomes strong when it reaches that portion of the basilar membrane that has a natural resonant frequency equal to the respective sound frequency. At this point the basilar membrane can vibrate back and forth with such great ease that the energy in the

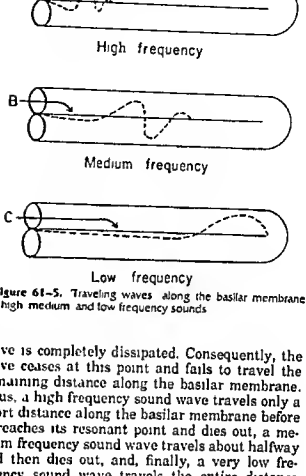


Figure 61–5. Traveling waves along the basilar membrane for high medium and low frequency sounds

wave is completely dissipated. Consequently, the wave ceases at this point and fails to travel the remaining distance along the basilar membrane. Thus, a high frequency sound wave travels only a short distance along the basilar membrane before it reaches its resonant point and dies out, a medium frequency sound wave travels about halfway and then dies out, and, finally, a very low frequency sound wave travels the entire distance along the membrane

Another feature of the traveling wave is that it travels fast along the initial portion of the basilar membrane but progressively more slowly as it goes farther and farther into the cochlea The cause of this is the high coefficient of elasticity of the basilar fibers near the stapes and a progressively decreasing coefficient farther along the membrane This rapid initial transmission of the wave allows the high frequency sounds to travel far enough into the cochlea to spread out and separate from each other on the basilar membrane Without this spread, all the high frequency waves would be bunched together within the first millimeter or so of the basilar membrane, and their frequencies could not be discriminated one from the other.

Amplitude Pattern of Vibration of the Basilar Membrane. The dashed curves of Figure 61–6A show the position of a sound wave on the basilar membrane when the stapes (a) is all the way inward, (b) has moved back to the neutral point, (c) is all the way outward, and (d) has moved back again to the neutral point but is moving inward. The shaded area around these different waves shows the maximum extent of vibration of the basilar membrane during a complete vibratory cycle. This is the amplitude pattern of vibration of the basilar membrane for this part[illegible] frequency.

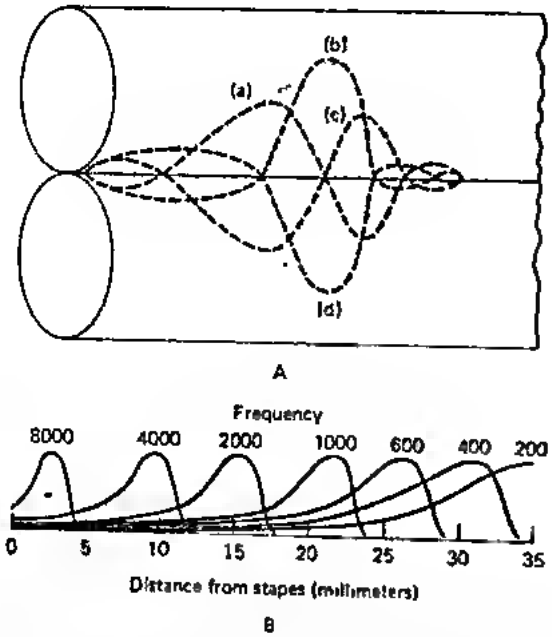


Figure 61–6. *A*, Amplitude pattern of vibration of the basilar membrane for a medium frequency sound. *B*, Amplitude patterns for sounds of all frequencies between 200 and 8000 per second showing the points of maximum amplitude (the resonance points) on the basilar membrane for the different frequencies.

Figure 61–6B shows the amplitude patterns of vibration for different frequencies, showing that the maximum amplitude for 8000 cycles occurs near the base of the cochlea, whereas that for frequencies less than 200 cycles per second is all the way at the tip of the basilar membrane near the helicotrema.

Note in Figure 61–6B that the basal end of the basilar membrane vibrates weakly for all frequencies. However, beyond the resonant area for each given frequency, the vibration of the basilar membrane cuts off sharply. The principal method by which sound frequencies, especially those above 200 cycles per second, are discriminated from each other is based on the "place" of maximum stimulation of the nerve fibers from the organ of Corti lying on the basilar membrane, as will be explained in the following section.

FUNCTION OF THE ORGAN OF CORTI

The organ of Corti, illustrated in Figures 61–2, 61–3, and 61–7, is the receptor organ that generates nerve impulses in response to vibration of the basilar membrane. Note that the organ of Corti lies on the surface of the basilar fibers and basilar membrane. The actual sensory receptors in the organ of Corti are two types of *hair cells*, a single row of *internal hair cells*, numbering about 3500 and measuring about 12 microns in diameter, and three to four rows of *external hair cells*, numbering about 20,000 and having diameters of only about 8 microns. The bases and sides of the hair cells synapse with a network of cochlear nerve endings.

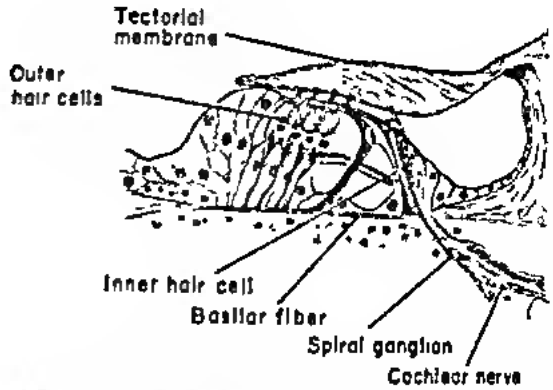


Figure 61–7. The organ of Corti, showing especially the i cells and the tectorial membrane against the projecting hairs

These lead to the *spiral ganglion of Corti*, whi lies in the modiolus (the center) of the cochle The spiral ganglion in turn sends axons into t *cochlear nerve* and thence into the central nervo system at the level of the upper medulla. T relationship of the organ of Corti to the spir ganglion and to the cochlear nerve is illustrat in Figure 61–2.

Excitation of the Hair Cells. Note in Figu 61–7 that minute hairs, or cilia, project upwa from the hair cells and either touch or are embe ded in the surface gel coating of the *tectori membrane* which lies above the cilia in the scal media. These hair cells are similar to the hair cel found in the macula and cristae ampullaris of th vestibular apparatus, which were discussed i Chapter 52. Bending of the hairs in one directio depolarizes the hair cells, and this in turn excite the nerve fibers synapsing with their bases.

Figure 61–8 illustrates the mechanism by whic vibration of the basilar membrane excites the hai endings. This shows that the upper ends of th hair cells are fixed tightly in a very rigid structur composed of a flat plate called the *reticular lamin* supported by triangular *rods of Corti* that in turn are attached tightly to the basilar fibers. There fore, the basilar fiber, the rods of Corti, and the reticular lamina all move as a rigid unit.

Upward movement of the basilar fiber rocks the reticular lamina upward and *inward*. Then, when the basilar membrane moves downward, the retic

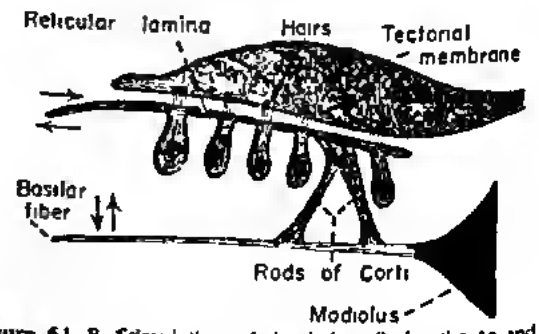


Figure 61–8. Stimulation of the hair cells by the to-and fro movement of the hairs in the tectorial membrane

ilar lamina rocks downward and *outward*. The nward and outward motion causes the hairs to hear back and forth against the tectorial membrane, thus exciting the cochlear nerve fibers whenever the basilar membrane vibrates.

Mechanism by Which the Hair Cells Excite the Nerve Fibers—Receptor Potentials. Back-and-forth bending of the hairs causes alternate changes in the electrical potential across the hair cell membrane. This alternating potential is the *receptor potential* of the hair cell, and it in turn stimulates the cochlear nerve endings that terminate on the hair cells.

When the basilar fiber bends toward the scala vestibuli (in the upward direction in Figure 61–8), the hair cell becomes depolarized, and it is this depolarization that generates action potentials in the nerve fiber. When the basilar fiber moves in the opposite direction, the hair cell becomes hyperpolarized, and the number of action potentials decreases.

The Endocochlear Potential. To explain even more fully the electrical potentials generated by the hair cells, we need to explain still another electrical phenomenon called the endocochlear potential: The scala media is filled with a fluid called *endolymph* in contradistinction to the *perilymph* present in the scala vestibuli and scala tympani. The scala vestibuli and scala tympani in most young children and in some adults communicate directly with the subarachnoid space around the brain, so that the perilymph is almost identical with cerebrospinal fluid. On the other hand, the endolymph that fills the scala media is an entirely different fluid probably secreted by the *stria vascularis*, a highly vascular area on the outer wall of the scala media. Endolymph contains a very high concentration of potassium and a very low concentration of sodium, which is exactly opposite to the perilymph.

An electrical potential of approximately +80 millivolts exists all the time between the endolymph and the perilymph, with positivity inside the scala media and negativity outside. This is called the *endocochlear potential*, and it is believed to be generated by continual secretion of positive potassium ions into the scala media by the stria vascularis.

The importance of the endocochlear potential is that the tops of the hair cells project through the reticular lamina into the endolymph of the scala media while perilymph bathes the lower bodies of the hair cells. Furthermore, the hair cells have a negative intracellular potential of −70 millivolts with respect to the perilymph, but −150 millivolts with respect to the endolymph at their upper surfaces where the hairs project into the endolymph. It is believed that this high electrical potential at the hair border of the cell greatly sensitizes the cell, thereby increasing its ability to respond to slight movement of the hairs.

DETERMINATION OF SOUND FREQUENCY—THE "PLACE" PRINCIPLE

From earlier discussions in this chapter it is already apparent that low frequency sounds cause maximal activation of the basilar membrane near the apex of the cochlea, sounds of high frequency activate the basilar membrane near the base of the cocblea, and intermediate frequencies activate the membrane at intermediate distances between these two extremes. Furthermore, there is spatial organization of the cochlear nerve fibers from the cochlea to the cochlear nuclei in the brain stem, with the fibers from each respective area of the basilar membrane terminating in a corresponding area of the cocblear nuclei. We sball see later that this spatial organization continues all the way up the brain stem to the cerebral cortex. The recording of alignals from the auditory tracts in the brain stem and from the auditory receptive fields in the cerebral cortex shows that specific neurons are activated by specific sound frequencies. Therefore, the major method used by the nervous system to detect different frequencies is to determine the position along the basilar membrane that is most stimulated. This is called the *place principle* for determination of frequency (or of sound "pitch").

Yet, referring again to Figure 61–6, one can see that the distal end of the basilar membrane at the helicotrema is stimulated by all sound frequencies below 200 cycles per second. Therefore, it has been difficult to understand from the place principle how one can differentiate between very low sound frequencies. It is postulated that these low frequencies are discriminated mainly by the so-called *frequency principle*. That is, the low frequency sounds cause volleys of impulses at the same low frequencies to be transmitted by the cochlear nerve into the cochlear nuclei. And it is believed that the cochlear nuclei then distinguish the different frequencies. In fact, destruction of the entire apical half of the cochlea, which destroys the basilar membrane where all the lower frequency sounds are normally detected, still does not completely eliminate the discrimination of low frequency sounds.

DETERMINATION OF LOUDNESS

Loudness is determined by the auditory system in at least three different ways: First, as the sound becomes louder, the amplitude of vibration of the basilar membrane and hair cells also increases so that the hair cells excite the nerve endings at more rapid rates. Second, as the amplitude of vibration increases, it causes more and more of the hair cells on the fringes of the resonating portion of the basilar membrane to become stimulated, thus causing *spatial summation* of impulses—that is, transmission through many nerve fibers, rather than through a few. Third, certain hair cells do not become stimulated until the vibration of the basilar membrane reaches a relatively high intensity, and it is believed that stimulation of these cells in some way apprises the nervous system that the sound is then very loud.

Detection of Changes in Loudness—The Power Law. It was pointed out in Chapter 48 that a person interprets changes in intensity of

sensory stimuli approximately in proportion to a power function of the actual intensity. In the case of sound, the interpreted sensation changes approximately in proportion to the cube root of the actual sound intensity. To express this another way, the ear can discriminate differences in sound intensity from the softest whisper to the loudest possible noise, representing an *approximate one trillion times* increase in sound energy or one million times increase in amplitude of movement of the basilar membrane. Yet the ear interprets this much difference in sound level as approximately a 10,000-fold change. Thus, the scale of intensity is greatly "compressed" by the sound perception mechanisms of the auditory system. This obviously allows a person to interpret differences in sound intensities over an extremely wide range, a far broader range than would be possible were it not for compression of the scale.

The Decibel Unit. Because of the extreme changes in sound intensities that the ear can detect and discriminate, sound intensities are usually expressed in terms of the logarithm of their actual intensities. A 10-fold increase in sound energy (or a $\sqrt{10}$-fold increase in sound pressure, because energy is proportional to the square of pressure) is called 1 *bel*, and 0.1 bel is called 1 *decibel*. One decibel represents an actual increase in sound energy of 1.26 times.

Another reason for using the decibel system in expressing changes in loudness is that, in the usual sound intensity range for communication, the ears can barely distinguish approximately a 1 decibel *change* in sound intensity.

The "Zero" Decibel Reference Level. The usual method for expressing the intensity of sound is to state the pressure difference between the peak of the sound compression wave and the trough of the wave. A pressure difference of *0.0002 dyne per square centimeter* is considered by audiologists to be unit intensity, and this is expressed as zero decibels when converted to the decibel scale. This sound level is also approximately the minimum that can be detected by the normal ear at the optimal frequency of 3000 cycles per second.

Threshold for Hearing Sound at Different Frequencies. Figure 61–9 shows the energy threshold at which sounds of different frequencies can barely be heard by the ear. This figure illustrates that a 3000 cycle per second sound can be heard even when its intensity is as low as 70 decibels below 1 dyne/cm² sound pressure level, which is one ten-millionth microwatt/cm². On the other hand, a 100 cycle per second sound can be detected only if its intensity is 10,000 times as great as this.

Frequency Range of Hearing. The frequencies of sound that a young person can hear, before aging has occurred in the ears, is generally stated to be between 20 and 20,000 cycles per second. However, referring again to Figure 61–9, we see that the sound range depends to a great extent on intensity. If the intensity is 60 decibels below the 1 dyne/cm² sound pressure level, the sound range is 500 to 5000 cycles per second, and only with intense sounds can the complete range of 20 – 20,000 cycles be achieved. In old age, the frequency range falls to 50 to 8000 cycles per second or less, as discussed later in the chapter.

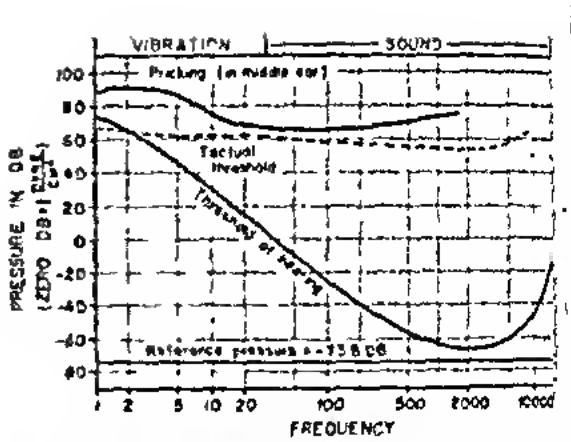


Figure 61–9. Relationship of the threshold of hearing and the threshold of somesthetic perception to the sound energy level at each sound frequency. (Modified from Stevens and Davis: Hearing. John Wiley & Sons.)

CENTRAL AUDITORY MECHANISMS

THE AUDITORY PATHWAY

Figure 61–10 illustrates the major auditory pathway. It shows that nerve fibers from the *spiral ganglion of Corti* enter the *dorsal* and *ventral cochlear nuclei* located in the upper part of the medulla. At this point, all the fibers synapse, and second order neurons pass mainly to the opposite side of the brain stem through the *trapezoid body* to the *superior olivary nucleus*. However, other second order fibers also pass ipsilaterally to the superior olivary nucleus on the same side. From the superior olivary nucleus the auditory pathway then passes upward through the *lateral lemniscus*, and many of the

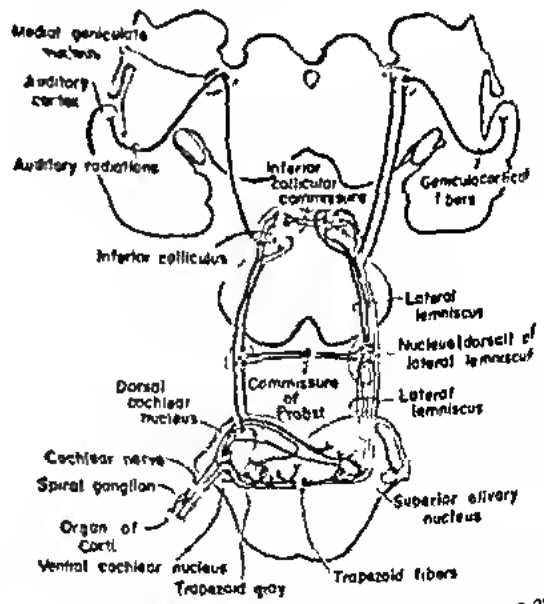


Figure 61–10. The auditory pathway. (Modified from Crosby, Humphrey, and Lauer: Correlative Anatomy of the Nervous System. The Macmillan Co.)

...a terminate in the *nucleus of the lateral lemniscus.* Many bypass this nucleus and pass on to the inferior colliculus where most terminate; a few pass on without terminating to higher levels. A few fibers cross from the nucleus of the lateral lemniscus through the *commissure of Probst* to the contralateral nucleus, and still other fibers cross through the *inferior collicular commissure* from one inferior colliculus to the other. From the inferior colliculus, the pathway passes to the *medial geniculate nucleus,* where all the fibers synapse. From here, the auditory pathway proceeds by way of the *auditory radiation* to the *auditory cortex* located mainly in the superior gyrus of the temporal lobe.

Several points of importance in relation to the auditory pathway should be noted. First, signals from both ears are transmitted through the pathways of both sides of the brain with only slight preponderance of transmission in the contralateral pathway. In at least three different places in the brain stem crossing-over occurs between the two pathways: (a) in the trapezoid body, (b) in the commissure of Probst, and (c) in the commissure connecting the two inferior colliculi.

Second, many collateral fibers from the auditory tracts pass directly into the *reticular activating system of the brain stem.* This system projects diffusely upward into the cerebral cortex and downward into the spinal cord. Other collaterals go to the *vermis of the cerebellum,* which is activated instantaneously in the event of a sudden noise.

Third, a high degree of spatial orientation is maintained in the fiber tracts from the cochlea all the way to the cortex. In fact, there are three different spatial representations of sound frequencies in the cochlear nuclei, two representations in the inferior colliculi, one precise representation for discrete sound frequencies in the auditory cortex, and at least five other less precise representations in the auditory cortex and auditory association areas.

Firing Rates at Different Levels of the Auditory Pathway. Single nerve fibers entering the cochlear nuclei from the eighth nerve can fire at rates up to at least 1000 per second, the rate being determined mainly by the loudness of the sound. At low sound frequencies, the nerve impulses are usually synchronized with the sound waves but they do not necessarily occur with every wave.

In the auditory tracts of the brain stem, the firing is usually no longer synchronized with the sound frequency except at sound frequencies below 200 cycles per second. And above the level of the inferior colliculi, even this synchronization is mainly lost. These findings demonstrate that the sound signals are not transmitted unchanged directly from the ear to the higher levels of the brain; instead, information from the sound signals begins to be dissected from the impulse traffic at levels as low as the cochlear nuclei. We will have more to say about this later, especially in relation to perception of direction from which sound comes.

Another significant feature of the auditory pathways is that low rates of impulse firing continue even in the absence of sound all the way from the cochlear nerve fibers to the auditory cortex. When the basilar membrane moves toward the scala vestibuli, the impulse traffic increases; and when the basilar membrane moves toward the scala tympani, the impulse traffic decreases. Thus, the presence of this background signal allows information to be transmitted from the basilar membrane when the membrane moves in either direction: positive information in one direction and negative information in the opposite direction. Were it not for the background signal, only the positive half of the information could be transmitted. This type of so-called "carrier wave" method for transmitting information is utilized in many parts of the brain, as has been discussed in several of the preceding chapters.

Discriminatory Functions of the Auditory Relay Nuclei. Cats and even monkeys can detect very low intensity sound even when the cerebral cortex is removed bilaterally, which indicates that the nuclei in the brain stem and thalamus can perform many auditory functions without the cerebral cortex. However, discrimination of tonal patterns, sound sequences, and direction from which the sound is coming are considerably impaired. In human beings, *bilateral* destruction of the cortical auditory centers may give a different picture; it is said to cause severe hearing loss but, unfortunately, this has not been studied adequately.

FUNCTION OF THE CEREBRAL CORTEX IN HEARING

The projection of the auditory pathway to the cerebral cortex is illustrated in Figure 61–11, which shows that the auditory cortex lies principally on the *supratemporal plane of the superior temporal gyrus* but also extends over the *lateral border of the temporal lobe,* over much of the *insular cortex,* and even into the most lateral portion of the *parietal operculum.*

Two separate areas are shown in Figure 61–11: the *primary auditory cortex* and the *auditory association cortex* (also called the *secondary auditory cortex*). The primary auditory cortex is directly excited by projections from the medial geniculate

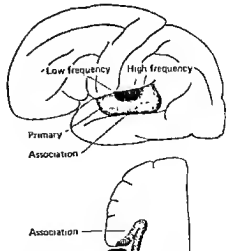


Figure 61–11. The auditory

body, while the auditory association areas are excited secondarily by impulses from the primary auditory cortex and by projections from thalamic association areas adjacent to the medial geniculate body.

Locus of Sound Frequency Perception in the Primary Auditory Cortex. At least six different *tonotopic maps* have been found in the primary auditory cortex and auditory association areas. In each of these maps, high frequency sounds excite neurons at one end of the map while low frequency sounds excite the neurons at the opposite end. In most, the low frequency sounds are located anteriorly, as shown in Figure 61–11, and the high frequency sounds posteriorly. However, this is not true for all the maps. The question that one must ask is why does the auditory cortex have so many different tonotopic maps? The answer is presumably that each of the separate areas dissects out some specific feature of the sounds. For instance, one of the large maps in the primary auditory cortex almost certainly discriminates the sound frequencies themselves and gives the person the psychic sensation of sound pitches. Another one of the maps probably is used to detect the direction from which the sound comes.

The frequency range to which each individual neuron in the auditory cortex responds is much narrower than that in the cochlear and brain stem relay nuclei. Referring back to Figure 61–6B, we note that the basilar membrane near the base of the cochlea is stimulated by all frequency sounds, and in the cochlear nuclei this same breadth of sound representation is found. Yet by the time the excitation has reached the cerebral cortex, most sound-responsive neurons respond only to a narrow range of frequencies rather than a broad range. Therefore, somewhere along the pathway, processing mechanisms "sharpen" the frequency response. It is believed that this sharpening effect is caused mainly by the phenomenon of lateral inhibition, which was discussed in Chapter 47 in relation to mechanisms for transmitting information in nerves. That is, stimulation of the cochlea at one frequency causes inhibition of signals caused by sound frequencies on either side of the stimulated frequency, this resulting from collateral fibers angling off the primary signal pathway and exerting inhibitory influences on adjacent pathways. The same effect has also been demonstrated to be important in sharpening patterns of somesthetic images, visual images, and other types of sensations.

A large share of the neurons in the auditory cortex, especially in the auditory association cortex, do not respond to specific sound frequencies in the ear. It is believed that these neurons "associate" different sound frequencies with each other or associate sound information with information from other sensory areas of the cortex. Indeed, the parietal portion of the auditory association cortex ...ly overlaps somatic sensory area II, which could provide easy opportunity for association of auditory information with somatic sensory information.

Discrimination of Sound "Patterns" by the Auditory Cortex. Complete bilateral removal of the auditory cortex does not prevent an animal from detecting sounds or reacting in a crude manner to the sounds. However, it does greatly reduce or sometimes even abolish its ability to discriminate different sound pitches and especially *patterns of sound*. For instance, an animal that has been trained to recognize a combination or sequence of tones, one following the other in a particular pattern, loses this ability when the auditory cortex is destroyed, and, furthermore, it cannot relearn this type of response. Therefore, the auditory cortex is important in the discrimination of *tonal* and *sequential sound patterns*.

In the human being, lesions affecting the auditory association areas but not affecting the primary auditory cortex do not decrease the person's full capability to hear and differentiate sound tones and to interpret at least simple patterns of sound. However, he or she will often be unable to interpret the *meaning* of the sound heard. For instance, lesions in the posterior portion of the superior temporal gyrus, which is Wernicke's area and is also part of the auditory association cortex, often make it impossible for the person to interpret the meanings of words even though he hears them perfectly well and can even repeat them. These functions of the auditory association areas and their relationship to the overall intellectual functions of the brain were discussed in detail in Chapter 54.

DISCRIMINATION OF DIRECTION FROM WHICH SOUND EMANATES

A person determines the direction from which sound emanates by two principal mechanisms: (1) by the time lag between the entry of sound into one ear and into the opposite ear and (2) by the difference between the intensities of the sounds in the two ears. The first mechanism functions best for frequencies below 3000 cycles per second, and the intensity mechanism operates best at higher frequencies because the head acts as a sound barrier at these frequencies. The time lag mechanism discriminates direction much more exactly than the intensity mechanism, for the time lag mechanism does not depend on extraneous factors but only on an exact interval of time between two acoustical signals. If a person is looking straight toward the sound, the sound reaches both ears at exactly the same instant, while, if the right ear is closer to the sound than the left ear, the sound signals from the right ear enter the brain ahead of those from the left ear.

Neural Mechanisms for Detecting Sound Direction. Destruction of the auditory cortex on both sides of the brain, in either human beings or lower

mammals, causes loss of almost all ability to detect the direction from which sound comes. Yet, the mechanism for this detection process begins in the superior olivary nuclei, even though it requires the neural pathways all the way from these nuclei to the cortex for interpretation of the signals. The mechanism is believed to be the following:

First, the superior olivary nucleus is divided into two sections, (a) the *medial superior olivary nucleus* and (b) the *lateral superior olivary nucleus*. The lateral nucleus is concerned with detecting the direction from which the sound is coming by the *difference in intensities of the sound* reaching the two ears, presumably by simply comparing the two intensities and sending an appropriate signal to the auditory cortex to estimate the direction

The *medial superior olivary nucleus*, on the other hand, has a very specific mechanism for *detecting the time-lag between acoustic signals entering the two ears*. This nucleus contains large numbers of neurons that have two major dendrites, one projecting to the right and the other to the left. The acoustical signal from the right ear impinges on the right dendrite and the signal from the left ear impinges on the left dendrite. The intensity of excitation of each of these neurons is highly sensitive to a specific time-lag between the two acoustical signals from the two ears. That is, the neurons near one border of the nucleus respond maximally to a short time-lag; whereas those near the opposite border respond to a very long time-lag; and those between, to intermediate time-lags. Thus, a spatial pattern of neuronal stimulation develops in the medial superior olivary nucleus, with sound from directly in front of the head stimulating one set of olivary neurons maximally and sounds from different side angles stimulating other sets of neurons maximally. This spatial orientation of signals is then transmitted all the way to the auditory cortex where sound direction is determined by the locus in the cortex that is stimulated maximally. It is believed that the signals for determining sound direction are transmitted through a different pathway and that this pathway terminates in the cerebral cortex in a different locus from the transmission pathway and termination locus for the tonal patterns of sound.

This mechanism for detection of sound direction indicates again how information in sensory signals is dissected out as the signals pass through different levels of neuronal activity. In this case, the "quality" of sound direction is separated from the "quality" of sound tones at the level of the superior olivary nuclei.

CENTRIFUGAL CONDUCTION OF IMPULSES FROM THE CENTRAL NERVOUS SYSTEM

Retrograde pathways have been demonstrated at each level of the nervous system from the auditory cortex to the cochlea. The final pathway is mainly from the superior olivary nucleus to the organ of Corti.

These retrograde fibers are inhibitory. Indeed, direct stimulation of discrete points in the olivary nucleus have been shown to inhibit specific areas of the organ of Corti, reducing their sound sensitivities as much as 15 to 20 decibels. One can readily understand how this could allow a person to direct attention to sounds of particular qualities while rejecting sounds of other qualities This is readily demonstrated when one listens to a single instrument in a symphony orchestra.

HEARING ABNORMALITIES

TYPES OF DEAFNESS

Deafness is usually divided into two types: first, that caused by impairment of the cochlea or auditory nerve, which is usually classed under the heading "nerve deafness," *and, second, that caused by impairment of the* middle ear mechanisms for transmitting sound into the cochlea, which is usually called "conduction deafness." Obviously, if either the cochlea or the auditory nerve is completely destroyed the person is permanently deaf. However, if the cochlea and nerve are still intact but the ossicular system has been destroyed or ankylosed ("frozen" in place by fibrosis or calcification), sound waves can still be conducted into the cochlea by means of bone conduction.

The Audiometer. To determine the nature of hearing disabilities, the audiometer is used. This is simply an earphone connected to an electronic oscillator capable of emitting pure tones ranging from low frequencies to high frequencies. Based on previous studies of normal persons, the instrument is calibrated so that the zero intensity level of sound at each frequency is the loudness that can barely be heard by the normal person. However, a calibrated volume control can increase or decrease the loudness of each tone above or below the zero level. If the loudness of a tone must be increased to 30 decibels above normal before it can be heard, the person is said to have a *hearing loss* of 30 decibels for that particular tone

In performing a hearing test using an audiometer, one tests approximately 8 to 10 frequencies covering the auditory spectrum, and the hearing loss is determined for each of these frequencies. Then the so-called "audiogram" is plotted as shown in Figures 61-12 and 61-13,

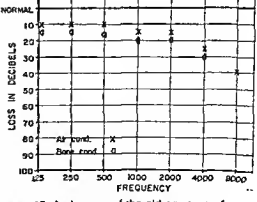


Figure 61-12. Audiogram of the old-age type of ness.

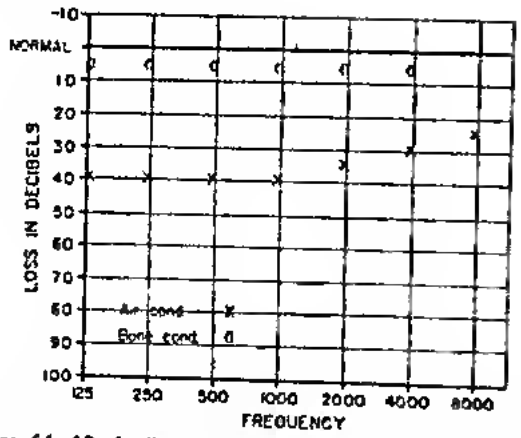


Figure 61–13. Audiogram of deafness resulting from middle ear sclerosis

depicting the hearing loss for each of the frequencies in the auditory spectrum

The audiometer, in addition to being equipped with an earphone for testing air conduction by the ear, is also equipped with an electronic vibrator for testing bone conduction from the mastoid process into the cochlea

The Audiogram in Nerve Deafness. In nerve deafness—this term including damage to the cochlea, to the auditory nerve, or to the central nervous system circuits from the ear—the person has decreased or total loss of ability to hear sound as tested by both the air conduction apparatus and the bone conduction apparatus An audiogram depicting partial nerve deafness is illustrated in Figure 61–12 In this figure the deafness is mainly for high frequency sound. Such deafness could be caused by damage to the base of the cochlea This type of deafness occurs to some extent in almost all older persons

Other patterns of nerve deafness frequently occur as follows. (1) deafness for low frequency sounds caused by excessive and prolonged exposure to very loud sounds (the rock band or the jet airplane engine) because low frequency sounds are usually louder and more damaging to the organ of Corti, and (2) deafness for all frequencies caused by drug sensitivity of the organ of Corti, especially sensitivity to some antibiotics such as streptomycin, kanamycin, and chloramphenicol.

The Audiogram in Conduction Deafness. A second and frequent type of deafness is that caused by fibrosis of the middle ear following repeated infection in the middle ear or fibrasis occurring in the hereditary disease called *otosclerosis* In this instanre the sound waves cannot be transmitted easily through the ossicles from the tympanic membrane to the oval window. Figure 61–13 illustrates an audiogram fram a person with "middle ear deafness" of this type In this case the bone conduction is essentially normal, but air conduction is greatly depressed at all frequencies, more so at the low frequencies In this type of deafness, the faceplate of the stapes frequently becomes "ankylosed" by bony overgrowth to the edges of the oval window. In this case, the person becomes totally deaf for air conduction, but can be made to hear again almost normally by removing the stapes and replacing it with a minute Teflon or metal prosthesis that transmits the sound from the incus to the oval dow

REFERENCES

Aitkin, L. M: Tonotopic organization at higher levels of the aud n pathway *Int Rev Physiol*, *10*:249 1976.

Aitkin, L. M *et al* Central neural mechanisms of hearing *In* Darnas Smith, I. (ed) Handbook of Physiology Sec 1, Vol III Bethesd American Physiological Society, 1984, p 675

Ballenger J J (ed) Diseases of the Nose, Throat, Ear, Head, and Neck Philadelphia, Lea & Febiger, 1985

Beagley, H A (ed): Auditory Investigation The Scientific and Technological Basis. New York, Oxford University Press, 1979

Becker W Atlas of Ear, Nose and Throat Diseases, Including Bronchoesophagology Philadelphia, W B Saunders Company, 1984

Bench, J, and Bamford, J (eds) Speech Hearing Tests and the Spoken Language of Hearing Impaired Children New York Academic Press 1979

Brugge, J F, and Geisler, C D: Auditory mechanisms of the lower brainstem. *Annu Rev Neurosci*, *1* 363, 1978

Busnel R, and Fish J F (eds) Animal Sonar Systems. New York Plenum Press, 1979

Carterette, E C, and Friedman, M P Hearing New York, Academic Press, 1978

Dallos, P Peripheral mechanisms of hearing *In* Darian-Smith, I. (ed) Handbook of Physiology Sec 1 Vol III Bethesda, American Physiological Society, 1984, p 595

Davis H Biophysics and physiology of the inner ear *Physiol Rev* 37 1, 1957

Evans, E F, and Wilson, J P (eds) Psychophysics and Physiology of Hearing New York, Academic Press, 1977

Green D M, and Wier C C Auditory perception *In* Darian-Smith, I (ed) Handbook of Physiology Sec 1, Vol III. Bethesda, American Physiological Society, 1984, p 557

Guth, P S, and Melamed, B Neurotransmission in the auditory system: A primer for pharmacologists. *Annu Rev Pharmacol Toxicol*, *22* 283, 1982

Harris, J D The Electrophysiology and Layout of the Auditory Nervous System Indianapolis The Bobbs-Merrill Co, 1974

Heasley, B E Auditory Processing Disorders and Remediation, 2nd Ed Springfield Ill Charles C Thomas, 1980

Hudspeth A J Mechanoelectrical transduction by hair cells in the acousticolateralis sensory system *Annu Rev Neurosci*, *6* 187, 1983

Imig, T J and Morel, A Organization of the thalamocortical auditory system in the cat *Annu Rev Neurosci*, *6* 95, 1983.

Kay, R H Hearing of modulation in sounds *Physiol Rev*, *62* 894, 1982

Kiang, N Y S Peripheral neural processing of auditory information. *In* Darian-Smith, I (ed) Handbook of Physiology Sec 1, Vol III Bethesda, American Physiological Society, 1984, p 639

Masterton, R B, and Imig T J Neural mechanisms of sound localization *Annu Rev Physiol*, *46* 275, 1984

Naunton, R F, and Fernandez C (eds.) Evoked Electrical Activity in the Auditory Nervous System New York, Academic Press, 1978

Rhode, W S Cochlear mechanics *Annu Rev Physiol*, *46* 231, 1984

Rosenblum, E H Fundamentals of Hearing for Health Professionals Boston, Little, Brown, 1979

Rubel, E W Ontogeny of auditory system function *Annu Rev Physiol*, *46* 213, 1984

Sachs, M B Neural coding of complex sounds. Speech *Annu Rev Physiol*, *46* 261, 1984

Sataloff, J, *et al* Hearing Loss 2nd Ed. Philadelphia, J B Lippincott Co, 1980

Scherich, O C H, and Schreiner, C (eds) Hearing Mechanisms and Speech New York, Springer Verlag, 1979

Schneiderman, C R. Basic Anatomy and Physiology in Speech and Hearing San Diego, College-Hill Press, 1984

Simmons, J A, and Kick S A Physiological mechanisms for spatial filtering and image enhancement in the sonar of bats *Annu Rev Physiol*, *46* 599, 1984

Singh, R. P Anatomy of Hearing and Speech. New York, Oxford University Press, 1980

Skinner, P H, and Shelton R L: Speech, Language, and Hearing Normal Processes and Disorders Reading, Mass, Addison Wesley Publishing Co, 1978

Stevens, S S: Hearing Its Psychology and Physiology New York. Acoustical Society of America, 1983

Van Hattum, R. J. Communication Disorders. New York, The Macmillan Co, 1980

Weiss, T F Relation of receptor potentials of cochlear hair cells to spike discharges of cochlear neurons *Annu Rev Physiol*, *46* 247, 1984

Wever, E G, and Lawrence, M Physiological Acoustics. Princeton, Princeton University Press, 1954

Woolsey, D N (ed) Cortical Sensory Organization Multiple Auditory Areas. Clifton, NJ, Humana Press, 1982

62

The Chemical Senses—Taste and Smell

HE SENSE OF TASTE

Taste is mainly a function of the *taste buds* in he mouth, but it is common experience that one's ense of smell also contributes strongly to taste erception. In addition, the texture of food, as letected by tactual senses of the mouth, and the resence of such elements in the food aa pepper, vhich stimulate pain endings, greatly condition he taste experience. The importance of taste lies n the fact that it allowa a person to select food in ccord with desirea and perhaps also in accord vitb the needs of the tissues for specific nutritive ubstances.

On the basia of paychological studies, there are enerally stated to be four *primary* sensations of aste: *sour, salty, sweet, and bitter.* Yet we know hat a person can perceive literally hundreds of lifferent tastes. These are all supposed to be comlinations of the four primary sensations in the ame manner that all the colors of the spectrum re combinations of three primary color sensations, s described in Chapter 59.

THE PRIMARY SENSATIONS OF TASTE

The Sour Taste. The sour taste is caused by acids, and the intensity of the taste sensation is pproximately proportional to the logarithm of the *hydrogen ion concentration.* That is, the more acidic the acid, the stronger becomes the sensation.

The Salty Taste. The salty taste is elicited by ionized salts. The quality of the taste varies somewhat from one salt to another because the salts also elicit other taste sensations besides saltiness. The cations of the salts are mainly responsible for the salty taste, but the anions also contribute to a lesser extent.

The Sweet Taste. The sweet taste is not caused by any single class of chemicals. A list of some of the types of chemicals that cause this taste includes sugars, glycols, alcohols, aldehydes, ketones, amides, esters, amino acids, sulfonic acids, halogenated acids, and inorganic salts of lead and beryllium. Note specifically that most of the substances that cause a sweet taste are organic chemicals. It is especially interesting that very slight changea in the chemical structure, such as addition of a simple radical, can often change the substance from sweet to bitter.

The third column of Table 62–1 shows the relative intensities of taste of certain substancea that cause the sweet taste. *Sucrose,* which is common table sugar, is considered to have an index of 1. Note that one of the substances has a aweet index 5000 times as great as that of aucrose. However, this extremely sweet substance, known as *P-4000,* is unfortunately extremely toxic and therefore cannot be used as a sweetening agent. *Saccharin,* on the other hand, is also more than 600 timea aa sweet as common table sugar, and since it is not toxic (except that it might be mildly carcinogenic), it can be used as a sweetening ogent.

The Bitter Taste. The bitter taste, like the sweet taste, is not caused by any single type of chemical agent, but, here again, the substances that give the bitter taste are almost entirely organic substances. Two particular classes of substances are especially likely to cause bitter tasto sensations: (1) long chain organic substances containing nitrogen and (2) alkaloids. The alkaloids include many of the drugs used in medicines such as quinine, caffeine, strychnine, and nicotine.

Some substances that at first taste sweet have a bitter aftertaste. This is true of saccharin, which makes this substance objectionable to some people. Some substances have a sweet taste on tbe front of the tongue, where taste buds with special sensitivity to the sweet taste are principally located, and a bitter taste on the back of the tongue, where taste buds more sensitive to the bitter taste are located.

The bitter taste, when it occurs in high intensity, usually causes the person or animal to reject the food. This is undoubtedly an important purposive function of the bitter taste sensation because many

Table 62–1 RELATIVE TASTE INDICES OF DIFFERENT SUBSTANCES

Sour Substances	Index	Bitter Substances	Index	Sweet Substances	Index	Salty Substances	Index
Hydrochloric acid	1	Quinine	1	Sucrose	1	NaCl	1
Formic acid	1.1	Brucine	11	1-propoxy-2-amino-4 nitrobenzene	5000	NaF	2
Chloracetic acid	0 9	Strychine	3 1	Saccharin	675	$CaCl_2$	1
Acetyllactic acid	0 85	Nicotine	1.3	Chloroform	40	NaBr	0 4
Lactic acid	0 85	Phenylthiourea	0 9	fructose	1 7	NaI	0.35
Tartaric acid	0 7	Caffeine	0 4	Alanine	1.3	LiCl	0 4
Malic acid	0 6	Veratrine	0 2	Glucose	0 8	NH_4Cl	2 5
Potassium H tartrate	0 58	Pilocarpine	0 16	Maltose	0 45	KCl	0 6
Acetic acid	0 55	Atropine	0 13	Galactose	0 32		
Citric acid	0 46	Cocaine	0 02	Lactose	0 3		
Carbonic acid	0 06	Morphine	0 02				

(From Derma. *Proc. Oklahoma Acad. Sci.*, 27 9, 1947, and Pfaffman, Handbook of Physiology Sec. I Vol. I. Baltimore, Williams & Wilkins, 1959, p 507)

of the deadly toxins found in poisonous plants are alkaloids, and these all cause intensely bitter taste.

Threshold for Taste

The threshold for stimulation of the sour taste by hydrochloric acid averages 0 0009 N; for stimulation of the salty taste by sodium chloride, 0.01 M; for the sweet taste by sucrose, 0 01 M, and for the bitter taste by quinine, 0.000008 M. Note especially how much more sensitive is the bitter taste sense than all the others, which would be expected since this sensation provides an important protective function.

Table 62–1 gives the relative taste indices (the reciprocals of the taste thresholds) of different substances In this table, the intensities of the four different primary sensations of taste are referred, respectively, to the intensities of taste of hydrochloric acid, quinine, sucrose, and sodium chloride, each of which is considered to have a taste index of 1

Taste Blindness. Many persons are taste blind for certain substances, especially for different types of thiourea compounds A substance used frequently by psychologists for demonstrating taste blindness is *phenylthiocarbamide*, for which approximately 15 to 30 per cent of all people exhibit taste blindness, the exact percentage depending on the method of testing and the concentration of the substance.

THE TASTE BUD AND ITS FUNCTION

Figure 62–1 illustrates a taste bud, which has a diameter of about 1/30 millimeter and a length of about 1/16 millimeter. The taste bud is composed of about 40 modified epithelial cells, some of which are supporting cells called *sustentacular cells* and others are *taste cells*. The taste cells are continually being replaced by mitotic division from the surrounding epithelial cells so that some are young cells and others are mature cells that lie toward the center of the bud and soon dissolute. The life span of each taste cell is about ten days in lower mammals but is unknown for the human being.

The outer tips of the taste cells are arranged around a minute *taste pore*, shown in Figure 62–1. From the tip of each cell, several *microvilli*, or *taste hairs*, about 2 to 3 microns in length and 0.1 to 0.2 micron in width, protrude outward into the taste pore to approach the cavity of the mouth. These microvilli are believed to provide the receptor surface for taste.

Interwoven among the taste cells is a branching terminal network of several *taste nerve fibers* that are stimulated by the taste cells. Some of these fibers invaginate into folds of the taste cell membranes, so that there is extremely intimate contact between the taste cells and the nerves.

An interesting feature of the taste buds is that they completely degenerate when the taste nerve fibers are destroyed. Then, if the taste fibers regrow to the epithelial surface of the mouth, the local epithelial cells regroup themselves to form new taste buds. This illustrates the important principle of "tropic" function of nerve fibers in certain parts of the body. The cause of the tropism is unknown, but it has been postulated to be a protein tropic factor secreted by the nerve endings.

Location of the Taste Buds. The taste buds are found on three of four different types of papillae of the tongue, as follows (1) A large number of taste buds are on the walls of the troughs that surround the circumvallate papillae, which form a V line toward the posterior of the tongue (2) Moderate numbers of taste buds are on the fungiform papillae over the front surface of the tongue (3) Moderate numbers are on the foliate papillae located in the folds along the lateral surfaces of the tongue Additional taste buds are located on the

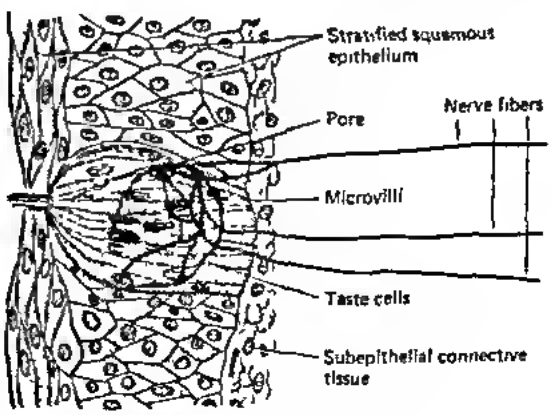


Figure 62–1. The taste bud.

palate and a few on the tonsillar pillars and at other points around the nasopharynx. Adults have approximately 10,000 taste buds, and children a few more. Beyond the age of 45 years many taste buds rapidly degenerate, causing the taste sensation to become progressively less critical.

Especially important in relation to taste is the tendency for taste buds subserving particular primary sensations of taste to be located in special areas. The sweet and salty tastes are located *principally* on the tip of the tongue, the sour taste on the two lateral sides of the tongue, and the bitter taste on the circumvallate papillae on the posterior tongue.

Specificity of Taste Buds for the Primary Taste Stimuli. Psychological tests using different types of taste stimuli carefully applied to individual taste buds, one at a time, have suggested that we have four distinctly different varieties of taste buds, each sensitive for only one type of taste. Yet, microelectrode studies from single taste buds while they are stimulated successively by the four different primary taste stimuli have shown that most of them can be excited by two, three, or even four of the primary taste stimuli, though usually with one or two of these predominating.

Regardless of the degree of specificity of the individual taste buds for different taste sensations, one can well understand that the hundreds of different types of tastes that we experience result from different quantitative degrees of stimulation of the four primary sensations of taste, as well as simultaneous stimulation of smell in the nose and tactile and pain nerve endings in the mouth.

Mechanism of Stimulation of Taste Buds. ***The Receptor Potential.*** The membrane of the taste cell, like that of other sensory receptor cells, is negatively charged on the inside with respect to the outside. Application of a taste substance to the taste hairs causes partial loss of this negative potential—that is, the taste cell is *depolarized*. The decrease in potential, within a wide range, is approximately proportional to the logarithm of concentration of the stimulating substance. This change in potential in the taste cell is the *receptor potential* for taste.

The mechanism by which the stimulating substance reacts with the taste hairs to initiate the receptor potential is unknown. It is believed by some physiologists that the substance is simply adsorbed to receptors on the membrane surface of the taste hair and that this adsorption changes the physical characteristics of the hair membrane. This in turn makes the taste cell more permeable to sodium ions and thus depolarizes the cell. The substance is gradually washed away from the taste hair by the saliva, thus removing the taste stimulus. Supposedly, the type or types of receptor in each taste hair determine the types of taste that will elicit responses.

Generation of Nerve Impulses by the Taste Bud. Most of the taste nerve fiber endings are encased by folds of the taste cell membranes. In some way not understood the receptor potentials of the taste cells generate impulses in the taste fibers, perhaps by secreting a chemical transmitter that in turn stimulates the nerve endings.

On first application of the taste stimulus, the rate of discharge of the nerve fibers rises to a peak in a small fraction of a second, but then it adapts within the next 2 seconds back to a lower steady level. Thus, a strong immediate signal is transmitted by the taste nerve, and a weaker continuous signal is transmitted as long as the taste bud is exposed to the taste stimulus.

TRANSMISSION OF TASTE SIGNALS INTO THE CENTRAL NERVOUS SYSTEM

Figure 62–2 illustrates the neuronal pathways for transmission of taste signals from the tongue and pharyngeal region into the central nervous system. Taste impulses from the anterior two thirds of the tongue pass first into the *fifth nerve* and then through the *chorda tympani* into the *facial nerve*, thence into the *tractus solitarius* in the brain stem. Taste sensations from the circumvallate papillae on the back of the tongue and from other posterior regions of the mouth are transmitted through the *glossopharyngeal nerve* also into the *tractus solitarius* but at a slightly lower level. Finally, a few taste signals are transmitted into the *tractus solitarius* from the base of the tongue and other parts of the pharyngeal region by way of the *vagus nerve*.

All taste fibers synapse in the *nuclei of the tractus solitarius* and send second order neurons to a small area of the *thalamus* located slightly medial to the thalamic terminations of the facial regions of the dorsal column–medial lemniscal system. From the thalamus, third order neurons are transmitted to the *lower tip of the postcentral gyrus in the parietal cortex* where it curls deep into the sylvian fissure. This lies in close association with, or even superimposed on, the tongue area of

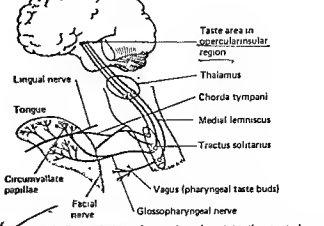


Figure 62–2. Transmission of taste impulses into the central nervous system.

somatic area I. Third order neurons also project to the nearby *opercular-insular area*, located deep in the sylvian fissure as well.

From this description of the taste pathways, it immediately becomes evident that they parallel closely the somatic pathways from the tongue.

Taste Reflexes. From the tractus solitarius a large number of impulses are transmitted directly into the *superior* and *inferior salivatory nuclei*, and these in turn transmit impulses to the submandibular, sublingual, and parotid glands to help control the secretion of saliva during the ingestion of food.

Adaptation of Taste. Everyone is familiar with the fact that taste sensations adapt rapidly, often with almost complete adaptation within a minute or so of continuous stimulation. Yet, from electrophysiological studies of taste nerve fibers, it seems that the taste buds themselves do not adapt enough to account for all or even most of the taste adaptation. The first burst of impulses from the taste bud allows one to detect extremely minute concentrations of taste substances, but normal concentrations of taste substances cause a weaker prolonged discharge of the taste fibers. Therefore, the progressive adaptation that occurs in the sensation of taste has been postulated to occur in the central nervous system itself, though the mechanism and site of this are not known. If this is true, it is a mechanism different from that of most other sensory systems, which adapt mainly at the receptors.

SPECIAL ATTRIBUTES OF THE TASTE SENSE

Affective Nature of Taste. *Pleasantness* and *unpleasantness* are called the "affective" attributes of a sensation. Figure 62–3 illustrates the affective effects of different types of taste at different concentrations of the stimulating substances, showing, strangely enough, that the sweet taste is likely to be unpleasant at a very low concentration but very pleasant at high concentrations. The other types of taste, on the other hand, are likely to be pleasant at low concentrations but exceedingly unpleasant at high concentrations. This is particularly true of the bitter taste.

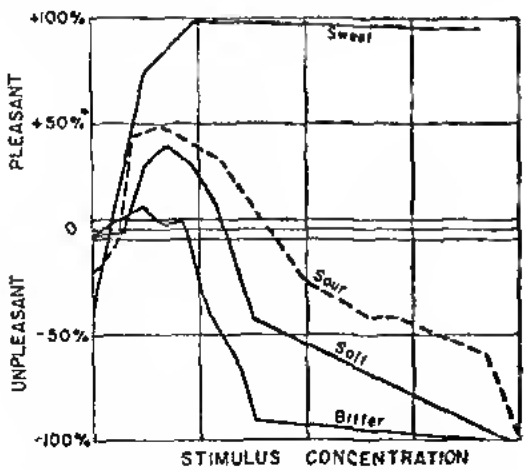


Figure 62–3. The affective nature of the different primary tions of taste at progressively increasing degrees of taste us. (From Engel, in Woodworth and Schlosberg, Experimentology. Holt, Rinehart & Winston, Inc. Reprinted with of CBS College Publishing.)

Importance of the Sense of Smell in Taste. Persons with severe colds frequently state that they have lost their sense of taste. However, on testing, the taste sensations are found to be completely normal. This illustrates that much of what we call taste is actually smell. Odors from the food can pass upward into the nasopharynx, often stimulating the olfactory system thousands of times as strongly as the taste system. For instance, if the olfactory system is intact, alcohol can be *"tasted" in 1/25,000 the concentration required when* the olfactory system is not intact.

Taste Preference and Control of the Diet. Taste preferences mean simply that an animal will choose certain types of food in preference to others, and it automatically uses this to help control the type of diet it eats. Furthermore, its taste preferences often change in accord with the needs of the body for certain specific substances. The following experimental studies illustrate this ability of an animal to choose food in accord with the need of its body. First, adrenalectomized animals automatically select drinking water with a high concentration of sodium chloride in preference to pure water, and this in many instances is sufficient to supply the needs of the body and prevent death as a result of salt depletion. Second, an animal injected with excessive amounts of insulin develops a depleted blood sugar, and it automatically chooses the sweetest food from among many samples. Third, parathyroidectomized animals automatically choose drinking water with a high concentration of calcium chloride.

These same phenomena are also observed in many instances of everyday life. For instance, the salt licks of the desert region are known to attract animals from far and wide, and even the human being rejects any food that has an unpleasant affective sensation, which certainly in many instances protects our bodies from undesirable substances.

The phenomenon of taste preference almost certainly results from some mechanism located in the central nervous system and not from a mechanism in the taste buds themselves, though it is true that the taste buds often do become sensitized to the needed nutrient. An important reason for believing taste preference to be mainly a central phenomenon is that previous experience with unpleasant or pleasant tastes plays a major role in determining one's different taste preferences. For instance, if a person becomes sick immediately after eating a particular type of food, the person generally develops a negative taste preference, or taste aversion, for that particular food thereafter; the same effect can be demonstrated in animals.

THE SENSE OF SMELL

Smell is the least understood sense. This results partly from the location of the olfactory membrane high in the nose where it is difficult to study and partly from the fact that the sense of smell is a subjective phenomenon that cannot be studied with ease in lower animals. Still another complicating problem is the fact that the sense of smell is almost rudimentary in the human being in comparison with that of some lower animals.

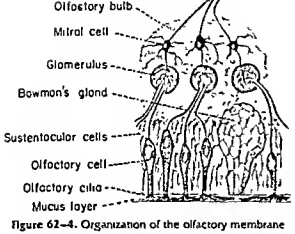


Figure 62–4. Organization of the olfactory membrane

THE OLFACTORY MEMBRANE

The olfactory membrane lies in the superior part of each nostril, as illustrated in Figure 62–5. Medially it folds downward over the surface of the septum, and laterally it folds over the superior turbinate and even over a small portion of the upper surface of the middle turbinate. In each nostril the olfactory membrane has a surface area of approximately 2.4 square centimeters.

The Olfactory Cells. The receptor cells for the smell sensation are the *olfactory cells*, which are actually bipolar nerve cells derived originally from the central nervous system itself. There are about 100 million of these cells in the olfactory epithelium interspersed among *sustentacular cells*, as shown in Figure 62–4. The mucosal end of the olfactory cell forms a knob from which 6 to 12 *olfactory hairs, or cilia*, 0.3 micron in diameter and 50 to 80 microns in length, project into the mucus that coats the inner surface of the nasal cavity. These projecting olfactory cilia form a dense mat in the mucus, and it is these cilia that react to odors in the air and then stimulate the olfactory cells, as discussed later. Spaced among the olfactory cells in the olfactory membrane are many small *glands of Bowman* that secrete mucus onto the surface of the olfactory membrane.

STIMULATION OF THE OLFACTORY CELLS

The Necessary Stimulus for Smell. We do not know what it takes chemically to stimulate the olfactory cells. Yet we do know the physical characteristics of the substances that cause olfactory stimulation: First, the substance must be volatile so that it can be sniffed into the nostrils. Second, it must be at least slightly water-soluble so that it can pass through the mucus to the olfactory cells. And, third, it must also be lipid-soluble, presumably because the olfactory cilia are composed principally of lipid materials.

Regardless of the basic mechanism by which the olfactory cells are stimulated, it is known that they become stimulated only when air blasts upward into the superior region of the nose. Therefore, smell occurs in cycles along with the inspirations, which indicates that the olfactory receptors respond in milliseconds to the volatile agents. Because smell intensity is exacerbated by blasting air through the upper reaches of the nose, the sensitivity of smell can be greatly increased by the well-known sniffing technique.

Receptor Potentials in Olfactory Cells. The olfactory cells react to olfactory stimuli in the same manner that most other sensory receptors react to their specific stimuli; that is, by depolarizing the cell and thus creating a receptor potential. This in turn initiates nerve impulses in the olfactory nerve fibers. An experiment that demonstrates this property of the olfactory receptors is the following: An electrode is placed on the surface of the olfactory membrane, and its electrical potential with respect to the remainder of the body is recorded. When an odorous substance is blown into the nostril, the surface potential becomes negative and remains negative as long as the odorous air continues to pass through the nostril. This electrical recording is called the *electro-olfactogram*, and it is believed to result from summation of receptor potentials developed in the receptor olfactory cells.

Over a wide range, both the amplitude of the electro-olfactogram and the rate of olfactory nerve impulses are approximately proportional to the logarithm of the stimulus strength, which illustrates that the olfactory receptors tend to obey principles of transduction similar to those of other sensory receptors.

Adaptation. The olfactory receptors adapt approximately 50 per cent in the first second or so after stimulation. Thereafter, they adapt very little and very slowly. Yet we all know from our own experience that smell sensations adapt almost to extinction within a minute or so after one enters a strongly odorous atmosphere. Since this psychological adaptation is far greater than the degree of adaptation of the receptors themselves, it is almost certain that the adaptation occurs in the central nervous system, which has also been postulated to be the locus of much of the adaptation for taste sensations. The neuronal mechanism for this adaptation could be the following: Large numbers of centrifugal nerve fibers pass from the olfactory regions of the brain backward along the olfactory tract and terminate on special inhibitory cells in the olfactory bulb, the *granule cells*. It is postulated that after the onset of an olfactory stimulus the central nervous system gradually develops a strong inhibitory signal to suppress transmission of the signal through the olfactory bulb.

Search for the Primary Sensations of Smell. Most physiologists are convinced that the many smell sensations are subserved by a few rather discrete primary sensations, in the same way that taste is subserved by sour, sweet, bitter, and salty sensations. But, thus far, only minor success has been achieved in classifying the primary sensations of smell. Yet, on the basis of psychological

tests and action potential studies from various points in the olfactory nerve pathways, it has been postulated that about seven different primary classes of olfactory stimulants preferentially excite separate olfactory cells. These classes of olfactory stimulants are characterized as follows:

1. Camphoraceous
2. Musky
3. Floral
4. Pepperminty
5. Ethereal
6. Pungent
7. Putrid

However, it is unlikely that this list actually represents the true primary sensations of smell, even though it does illustrate the results of one of the many attempts to classify them. Indeed, several clues in recent years have indicated that there may be as many as *50* or more primary sensations of smell—a marked contrast to only *three* primary sensations of color detected by the eyes and only *four* primary sensations of taste detected by the tongue. For instance, persons have been found who have *odor blindness* for single substances; and such discrete odor blindness has been identified for more than 50 different substances. Since it is presumed that odor blindness for each substance represents a lack of the appropriate receptor cell for that substance, it is postulated that the sense of smell might be subserved by 50 or more primary smell sensations.

Mechanism for Excitation of the Olfactory Cells. The basic mechanism by which the olfactory cells are stimulated is not known, but two principal theories have been proposed. One of these is a *physical theory* which suggests that it is the physical shapes of the odor molecules that determine which olfactory cells will be stimulated, but the evidence for this theory is weak. The second theory, a *chemical theory*, is much more likely to be correct. This theory assumes that the odor molecules bind chemically to protein *receptors* in the membranes of the olfactory cilia. The type of receptor in each olfactory cell determines the type of stimulant that will excite the cell. The reaction between the stimulant and the receptor supposedly increases the permeability of the olfactory ciliary membrane, and this in turn creates the receptor potential in the olfactory cell that generates impulses in the olfactory nerve fibers.

Affective Nature of Smell. Smell, equally as much as taste, has the affective qualities of either pleasantness or unpleasantness. Because of this, smell is as important as, if not more important than, taste in the selection of food. Indeed, a person who has previously eaten food that has disagreed with him is often nauseated by even the smell of that same type of food on a second occasion. Other types of odors that have proved to be unpleasant in the past may also provoke a disagreeable feeling; on the other hand, perfume of the right quality can wreak havoc with masculine emotions. In addition, in some lower animals odors are the primary excitant of sexual drive.

Threshold for Smell. One of the principal characteristics of smell is the minute quantity of the stimulating substance in the air often required to effect a smell sensation. For instance, the substance *methyl mercaptan* can be smelled when only 1/25,000,000,000 milligram is present in each milliliter of air. Because of this low threshold, this substance is mixed with natural gas to give the gas an odor that can be detected when it leaks from a gas pipe.

Measurement of Smell Threshold. One of the problems in studying smell has been difficulty in obtaining accurate measurements of the threshold stimulus required to induce smell. The simplest technique is simply to allow a person to sniff different substances in the usual manner of smelling. Indeed, some investigators feel that this is as satisfactory as almost any other procedure. However, to eliminate variations from person to person, more objective methods have been developed. One of these has been to place a box containing the volatilized agent over the subject's head. Appropriate precautions are taken to exclude odors from the person's own body. The person is allowed to breathe naturally, but the volatilized agent is distributed evenly in the air that is breathed.

Gradations of Smell Intensities. Though the threshold concentrations of substances that evoke smell are extremely slight, concentrations only 10 to 50 times above the threshold values often evoke maximum intensity of smell. This is in contrast to most other sensory systems of the body, in which the ranges of detection are tremendous—for instance, 500,000 to 1 in the case of the eyes and 1,000,000,000,000 to 1 in the case of the ears. This perhaps can be explained by the fact that smell is concerned more with detecting the presence or absence of odors than with quantitative detection of their intensities.

TRANSMISSION OF SMELL SIGNALS INTO THE CENTRAL NERVOUS SYSTEM

The function of the central nervous system in olfaction is almost as vague as the function of the peripheral receptors. However, Figures 62–4 and 62–5 illustrate the general plan for transmission of olfactory signals into the central nervous sys-

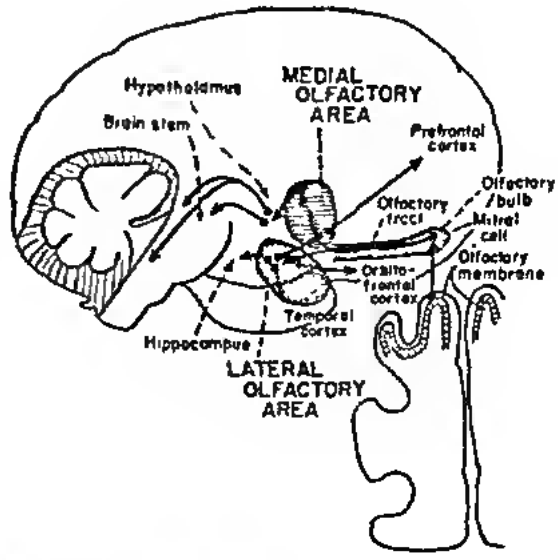


Figure 62–5. Neural connections of the olfactory system.

tem. Figure 62–4 shows a number of separate *olfactory cells* sending axons into the *olfactory bulb* to end on *dendrites from mitral cells* in a structure called the *glomerulus*. Approximately 25,000 axons enter each glomerulus and synapse with about 25 mitral cells that in turn send signals into the brain. Also in the glomerulus are about 60 smaller cells, the *tufted cells*, that send signals to the brain as well—perhaps to the older portions of the olfactory nervous system.

Figure 62–5 shows the major pathways for transmission of olfactory signals from the mitral and tufted cells into the brain. The fibers from the cells travel through the olfactory tract and terminate either primarily or after relay neurons in two principal areas of the brain called the *medial olfactory area* and the *lateral olfactory area*, respectively. The medial olfactory area is composed of a group of nuclei located in the midportion of the brain anterior to the hypothalamus. This group includes the *olfactory nucleus*, the *olfactory tubercle*, parts of the *hypothalamus*, and other adjacent areas.

The lateral olfactory area is composed mainly of the *prepyriform* and *pyriform cortex* and part of the *amygdaloid nuclei*.

Secondary olfactory tracts pass from both the medial and lateral olfactory areas into many other portions of the *limbic system* and into associated regions of the *thalamus* and *brain stem nuclei*.

The lateral olfactory area, especially part of the *amygdala* and the overlying cortex, the *pyriform* and *prepyriform* areas, is believed to be responsible for the more complex aspects of olfaction, such as association of olfactory sensations with somatic, visual, tactile, and other types of sensation. These regions of the brain, operating in association with other areas of the cortex located more anteriorly in the *uncus*, *orbitofrontal cortex*, and *frontal cortex*, are probably responsible for a person's specific appetites for certain types of foods. For instance, it might be in these areas that one develops special likes or dislikes for certain foods depending upon previous experience with the foods.

The *pyriform area* of the temporal cortex is frequently considered to be the *primary olfactory cortex*. It is particularly interesting that this is the only part of the entire cerebral cortex that receives primary sensory signals that are not relayed through the thalamus.

Complete removal of the lateral olfactory area hardly affects the primitive responses to olfaction, such as licking the lips, salivation, and other feeding responses caused by the smell of food or such as the various emotions associated with smell. On the other hand, its removal does abolish the more complicated conditioned reflexes that depend on olfactory stimuli.

Centrifugal Control of Activity in the Olfactory Bulb by the Central Nervous System. Many nerve fibers originating in the olfactory portions of the brain pass in the backward direction in the olfactory tract to the olfactory bulb, that is, "centrifugally" from the brain to the periphery. These terminate on the very large number of small *granule cells* located in the center of the bulb. These in turn send short, inhibitory *dendrites* to the mitral and tufted cells. As discussed earlier, it is believed that this inhibitory feedback to the olfactory bulb might be a method for central adaptation of olfaction to the different odors. It might also be a means for helping sharpen one's specific capability of distinguishing one odor from another.

Electrical Activity in the Olfactory Nerves and Tracts. The olfactory receptors in the olfactory epithelium, even normally, send impulses at a slow rate of 1 to 3 pulses per second. Most odors increase the impulse traffic, but a few actually inhibit it. A strong olfactory signal can increase the impulse rate to a maximum of about 20 per second.

Electrophysiological studies show that the mitral and tufted cells are also continually active. Superimposed on this background are increases or decreases in impulse traffic caused by different odors. Thus, the olfactory stimuli *modulate* the frequency of impulses in the olfactory system and in this way transmit the olfactory information.

REFERENCES

Alberts, J R. Producing and interpreting experimental olfactory deficits *Physiol Behav*, *12* 657, 1974

Dastoli, F R. Taste receptor proteins *Life Sci*, *14* 1417, 1974

Denton, D A. Salt appetite. *In* Code, C F, and Heidel, W (eds.). Handbook of Physiology Sec 6, Vol 1 Baltimore, Williams & Wilkins, 1967, p 433

Douek, E The Sense of Smell and Its Abnormalities New York, *Churchill Livingstone*, 1974

Hodgson, E S Taste receptors. *Sci Am*, *204* 135, 1961

Kare, M R., and Maller, O The Chemical Sense and Nutrition. Baltimore, The Johns Hopkins Press, 1967

Kashara, Y (ed) Proceedings of the Seventeenth Japanese Symposium on Taste and Smell Arlington, VA, IRL Press, 1984

Lat, J Self selection of dietary components *In* Code, C F, and Heidel, W (eds) Handbook of Physiology Sec. 6, Vol 1 Baltimore, Williams & Wilkins, 1967, p 367

McBurney, D H Taste and olfaction. Sensory discrimination. *In* Darian Smith, I (ed) Handbook of Physiology. Sec 1, Vol III. Bethesda, American Physiological Society, 1984, p 1067

Moulton, D G, and Beidler, L M Structure and function in the peripheral olfactory system *Physiol Rev*, *47* 1, 1967

Norgren, R Central neural mechanisms of taste *In* Darian-Smith, I. (ed) Handbook of Physiology. Sec. 1, Vol. III. Bethesda, American Physiological Society, 1984, p 1037

Oakley, B, and Benjamin, R M Neural mechanisms of taste. *Physiol. Rev*, *66* 173, 1966

Ohloff, G, and Thomas, A F (eds) Gustation and Olfaction. New York, Academic Press, 1971

Shepherd, G M. The olfactory bulb· A simple system in the mammalian brain *In* Brookhart, J M., and Mountcastle, V B (eds) Handbook of Physiology. Sec. 1, Vol 1. Baltimore, Williams & Wilkins, 1977, p 945

Takagi, S F. The olfactory nervous system of the Old World monkey. *Jpn. J. Physiol.*, *34* 561, 1984

Todd, J H. The chemical language of fishes *Sci. Am.*, *224* 98, 1971.

Wenzel, B. M., and Sieck, M. H. Olfaction. *Annu. Rev Physiol*, *28* 381, 1966.

Zotterman, Y· Olfaction and Taste. New York, The Macmillan Co, 1963

XI

THE GASTROINTESTINAL TRACT

63 ■ Movement of Food Through the Alimentary Tract

64 ■ Secretory Functions of the Alimentary Tract

65 ■ Digestion and Absorption in the Gastrointestinal Tract

66 ■ Physiology of Gastrointestinal Disorders

63

Movement of Food Through the Alimentary Tract

The primary function of the alimentary tract is to provide the body with a coatinual supply of water, electrolytes, and nutrients, but before this can be achieved food must be moved along the alimentary tract at an appropriate rate for the digestive and absorptive functions to take place. Therefore, discussion of the alimentary system is presented in three different phases in the next three chapters: (1) movement of food through the alimentary tract, (2) secretion of the digestive juices, and (3) absorption of the digested foods, water, and the various electrolytes.

Figure 63–1 illustrates the entire alimentary tract, showing major anatomical differences between its parts. Each part is adapted for apecific functions, such as: (1) simple passage of food from one point to another, as in the esophagus, (2) atorage of food in the body of the atomach or of fecal matter in the descending colon, (3) digestion of food in the stomach, duodenum, jejunum, and ileum, and (4) absorption of the digestive end-products and fluids in the entire small intestine and proximal half of the colon. One of the most important features of the gastrointestinal tract that is discussed in the present chapter is the myriad of autoregulatory processes in the gut that keeps the food moving at an appropriate pace—slow enough for digestion and absorption to toke place but fast enough to provide the nutrients needed by the body.

GENERAL PRINCIPLES OF GASTROINTESTINAL MOTILITY

CHARACTERISTICS OF THE GASTROINTESTINAL WALL

Figure 63–2 illustrates a typical section of the intestinal wall, showing the following layers from the outer surface inward: (1) the *serosa*, (2) a *gitudinal muscle layer*, (3) a *circular muscle* ., (4) the *submucosa*, and (5) the *mucosa*. In addition, a sparse layer of smooth muscle fib the *muscularis mucosae*, lies in the deeper lay of the mucosa. The motor functions of the gut performed by the different layers of smooth mus

The general characteristics of smooth mus and its function were discussed in Chapter However, the specific characteristics of smo muscle in the gut are the following:

The Gnstrointestinal Smooth Muscle—Function as a Syncytium. The individual smo muscle fibers in the gastrointestinal tract are tween 200 and 500 microns in length and 2 10 microns in diameter, and they are arranged bundles of as many as a thousand parallel fib

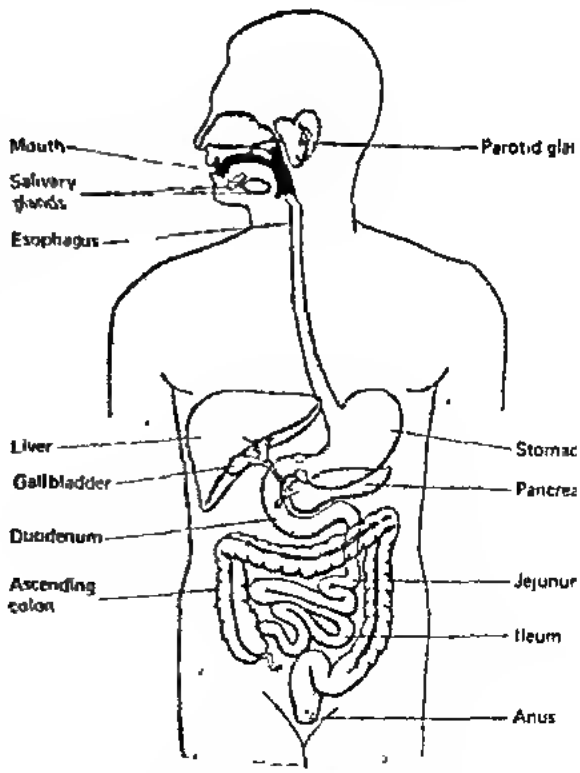


Figure 63–1. The alimentary tract.

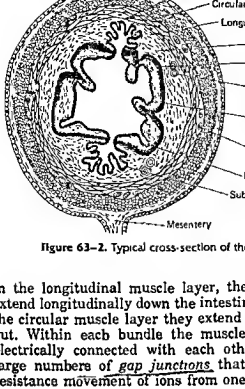


Figure 63–2. Typical cross-section of the gut.

In the longitudinal muscle layer, these bundles extend longitudinally down the intestinal tract; in the circular muscle layer they extend around the gut. Within each bundle the muscle fibers are electrically connected with each other through large numbers of *gap junctions* that allow low resistance movement of ions from one cell to the next. Therefore, electrical signals can travel readily from one fiber to the next.

It is the *bundles* of muscle fibers, not the individuel fibers, that conduct action potentials in the gastrointestinal musculature because single fibers cannot generate sufficient electrical current alone to elicit an action potential. At least 200 to 300 muscle fibers operating in parallel are required for this to occur.

Each bundle of smooth muscle fibers is separated from the next by loose connective tissue, but the bundles fuse with each other at many points so that in reality each muscle layer represents a branching latticework of smooth muscle bundles. Therefore, each muscle layer functions as a *syncytium*—that is, when an electrical signal is elicited anywhere within the muscle mass, it generally travels in all directions in the muscle. However, the distance that it travels depends upon the excitability of the muscle; sometimes it stops after only a few millimeters and at other times travels many centimeters or even the entire length of the intestinal tract. Also, a few connections exist between the longitudinal and circular muscle layers so that excitation of one of these layers usually excites the other as well.

Electrical Activity of Gastrointestinal Smooth Muscle

The smooth muscle of the gastrointestinal tract undergoes almost continual electrical activity. This activity is sometimes very bizarre but tends to have two basic types of electrical waves: (1) *slow waves* and (2) *spikes*, both of which are illustrated in Figure 63–3. In addition, the voltage of the resting membrane potential of the gastrointestinal smooth muscle can change to different levels without the generation of waves, and this too can have important effects in controlling motor activity of the gastrointestinal tract.

The Slow Waves—the "Basic Electrical Rhythm" (BER). Most gastrointestinal contractions occur rhythmically, and this rhythm is determined almost entirely by the frequency of the so-called "slow waves" in the smooth muscle membrane potential. These waves, illustrated in Figure 63–3, are called the "basic electrical rhythm" of the gut. The slow waves are not action potentials. Instead, they are slow, undulating changes in the resting membrane potential. Their intensity usually varies between 5 and 15 millivolts, and their frequency ranges in different parts of the human gastrointestinal tract between 3 and 12 per minute: about 3 in the body of the stomach, as much as 12 in the duodenum, and changing to about 8 or 9 in the terminal ileum. Therefore, the rhythm of contraction of the body of the stomach usually is about 3 per minute, of the duodenum about 12 per minute, and the ileum 8 to 9 per minute.

The cause of the slow waves is not known. However, as discussed in Chapter 12, it is believed that they might be caused by a slow undulation of the activity of the sodium-potassium pump.

The slow waves themselves *do not directly cause* muscle contraction. However, they do control the appearance of intermittent spike potentials, and the spike potentials in turn actually cause the muscle contraction.

The Spike Potentials. The spike potentials are true action potentials. They occur automatically when the resting membrane potential of the gastrointestinal smooth muscle rises above about −40 millivolts (the normal resting membrane potential is between −50 and −65 millivolts). Thus, note in Figure 63–3 that each time the peaks of the slow waves rise temporarily above the −40 milli-

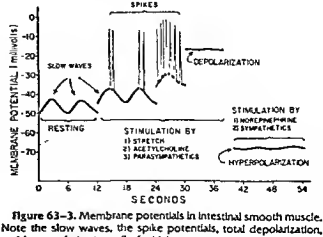


Figure 63–3. Membrane potentials in intestinal smooth muscle. Note the slow waves, the spike potentials, total depolarization, and hyperpolarization, all of which occur under different physiological conditions of the intestine.

volt level, spike potentials appear. And the higher the slow wave potential rises above this level, the greater the frequency of the spike potentials, usually ranging between 1 and 10 spikes per second.

The spike potentials last 10 to 40 times as long in the gastrointestinal muscle as the action potentials in most nerve fibers, each lasting as long as 10 to 20 milliseconds. Another very important difference between the action potentials of the gastrointestinal smooth muscle and those of nerve fibers is the manner in which they are generated. In nerve fibers, the action potentials are caused almost entirely by rapid entry of sodium ions through sodium channels to the interior of the fibers In the gastrointestinal smooth muscle, the channels responsible for the action potentials are quite different; they allow especially large amounts of calcium ions to enter along with smaller amounts of sodium ions and therefore are called *calcium-sodium channels*. These channels are much slower to open and also to close than are the rapid sodium channels, which accounts for the longer duration of the action potentials Also, the movement of larga amounts of calcium ions to the interior of the muscla fiber during the action potential is especially important in causing the intestinal smooth muscle to contract, as we shall discuss shortly

Even though the spike potentials are true action potentials, they usually do not propagate more than a few millimeters in the intestinal musculature Instead, it is mainly the slow wave potentials that are propagated, and the spike potentials then occur at the peaks of the slow waves in each local segment of tha intestinal tract. This is the reason why the slow waves rather than the spiko potentials establish the basic electrical rhythm.

Changes in the Voltage of the Resting Membrane Potential. In addition to the slow waves and spike potentials, the voltage level of the resting membrane potential can also change Under normal conditions the resting membrane potential averages about −56 millivolts, but multiple factors can change this level When the potential becomes more positive, which is called *depolarization* of the membrane, the muscle fiber becomes more excitable When the potential becomes more negative, which is called *hyperpolarization*, the fiber becomes less excitable Some factors that depolarize the membrane, that is, that make it more excitable are (1) stretch of the muscle, (2) stimulation by acetylcholine, (3) stimulation by the parasympathetic nerves, and (4) stimulation by several specific gastrointestinal hormones. Important factors that decrease the membrane potential to more negative levels—that is, hyperpolarize the membrane and make the muscle fiber less excitable—are (1) the effect of norepinephrine or epinephrine on the muscle membrane, and (2) stimulation of the sympathetic nerves that secrete norepinephrine at their endings.

Calcium Ions and Muscle Contraction. Muscle contraction occurs in response to the entry of calcium into the muscle fiber. As was explained in Chapter 12, it is believed that in smooth muscle the calcium ions activate the myosin filaments in the fiber, causing attractive forces between these and the actin filaments and thereby causing the muscle to contract.

The slow waves do not cause calcium ions to enter the smooth muscle fiber but instead only sodium ions Therefore, the slow waves by themselves cause no contraction. Instead, it is during the spike potentials at the peaks of the slow waves that large quantities of calcium ions do enter the fibers and cause the contractions.

Tonic Contraction of Gastrointestinal Smooth Muscle. The smooth muscle of the gastrointestinal tract usually exhibits *tonic contraction* as well as rhythmical contractions. Tonic contraction is continuous, not associated with the basic electrical rhythm of the slow waves but instead often lasting minute after minute or even hour after hour. The tonic contraction occasionally increases or decreases in intensity, but nevertheless continues Tonic contraction is sometimes caused by repetitive series of spike potentials—the greater the frequency the greater the degree of contraction However, at other times tonic contraction is caused by hormones or other factors that bring about continuous depolarization of the smooth muscle membrane without causing action potentials And still a third cause of tonic contraction is continuous entry of calcium into the interior of the cell brought about in ways not at all associated with changes in the membrane potential Unfortunately, the details of these mechanisms are still very unclear.

INNERVATION OF THE GUT—THE ENTERIC NERVOUS SYSTEM

The gastreintestinal tract has an intrinsic nervous system of its own, called the *enteric nervous system*, that begins in the esophagus and extends all the way to the anus. This system controls most gastrointestinal functions, especially gastrointestinal movements and secretion. On the other hand, both parasympathetic and sympathetic nervous signals to the gastrointestinal tract from the brain can strongly alter the degree of activity of this enteric nervous system.

The Myenteric Plexus and the Submucosal Plexus. The enteric nervous system is composed principally of two layers of neurons and appropriate connecting fibers· the outer layer, called the *myenteric plexus or Auerbach's plexus*, lies between the longitudinal and circular muscular layers; the inner layer, called the *submucosal plexus* or *Meissner's plexus*, lies in the submucosa. The myenteric plexus controls mainly the *gastrointestinal movements*, whereas the submucosal plexus is impor-

tant in controlling *secretion* and *blood flow* and also subserves many *sensory functions*, receiving signals principally from the gut epithelium and from stretch receptors in the gut wall.

Most often, stimulation of the myenteric plexus increases the motor activity of the gut, causing four principal effects: (1) increased tonic contraction, or "tone," of the gut wall, (2) increased intensity of the rhythmic contractions, (3) slightly increased rate of rhythmic contraction, and (4) increased velocity of conduction of excitatory waves along the gut wall. On the other hand, some myenteric plexus fibers are inhibitory rather than excitatory; these fibers secrete an inhibitory transmitter, possibly VIP (vasoactive intestinal polypeptide) or some other peptide. A large share of the excitatory fibers are *cholinergic*—that is, they secrete *acetylcholine*, though some secrete other excitatory transmitters. Transmitters that have been identified in the enteric plexuses, though their functions are mainly unknown, include *ATP*, *substance P*, *enkephalin*, *somatostatin*, *serotonin*, *bombesin*, and *neurotensin*.

Autonomic Control of the Gastrointestinal Tract

The gastrointestinal tract receives extensive parasympathetic and sympathetic innervation that is capable of altering the overall activity of the entire gut or of specific parts of it.

Parasympathetic Innervation. The parasympathetic supply to the gut is divided into *cranial* and *sacral divisions*, which were discussed in Chapter 57. Except for a few parasympathetic fibers to the mouth and pharyngeal regions of the alimentary tract, the cranial parasympathetics are transmitted almost entirely in the *vagus nerves*. These fibers provide extensive innervation to the esophagus, stomach, pancreas, and first half of the large intestine (but rather little innervation to the small intestine). The sacral parasympathetics originate in the second, third, and fourth sacral segments of the spinal cord and pass through the *pelvic nerves* to the distal half of the large intestine. The sigmoidal, rectal, and anal regions of the large intestine are considerably better supplied with parasympathetic fibers than are the other portions. These fibers function especially in the defecation reflexes, which are discussed later in the chapter.

The postganglionic neurons of the parasympathetic system are mainly located in the myenteric and submucosal plexuses, so that stimulation of the parasympathetic nerves causes a general increase in activity of the entire enteric nervous system. This in turn enhances the activity of most gastrointestinal functions, but not all, for some of the enteric neurons are inhibitory and therefore inhibit certain of the functions.

Sympathetic Innervation. The sympathetic fibers to the gastrointestinal tract originate in the spinal cord between the segments T-8 and L-2. The preganglionic fibers, after leaving the cord, enter the sympathetic chains and pass through the chains to outlying ganglia, such as the *celiac ganglion* and various *mesenteric ganglia*. Here, the postganglionic neuron bodies are located, and postganglionic fibers spread from them along with the blood vessels to all parts of the gut, terminating principally on neurons in the enteric nervous system. The sympathetics innervate essentially all portions of the gastrointestinal tract rather than being more extensively supplied to the most orad and most analward portions as is true of the parasympathetics. The sympathetic nerve endings secrete *norepinephrine*.

In general, stimulation of the sympathetic nervous system inhibits activity in the gastrointestinal tract, causing effects essentially opposite to those of the parasympathetic system. It exerts its effects in two different ways: (1) to a slight extent by direct effect of the norepinephrine on the smooth muscle to inhibit this (except the muscularis mucosa, which it excites), and (2) to a major extent by an inhibitory effect of the norepinephrine on the neurons of the enteric nervous system. Thus, strong stimulation of the sympathetic system can totally block movement of food through the gastrointestinal tract.

Afferent Nerve Fibers from the Gut. Many afferent nerve fibers arise in the gut. Some of these have their cell bodies in the enteric nervous system itself. These nerves can be stimulated by (a) irritation of the gut mucosa, (b) excessive distension of the gut, or (c) the presence of specific chemical substances in the gut. Signals transmitted through these fibers can cause excitation or, under some conditions, inhibition of intestinal movements or intestinal secretion.

In addition to the afferent fibers that terminate in the enteric nervous system, still two other types of afferent fibers are associated with this system. One of these has its cell bodies in the enteric nervous system but sends its axons through the extrinsic nerves to terminate in the *prevertebral sympathetic ganglia*, that is, in the *celiac*, *mesenteric*, and *hypogastric ganglia*. The other type of afferent fibers has its cell bodies in the dorsal root ganglia of the spinal cord or in the cranial nerve ganglia, these fibers transmit their signals directly into the spinal cord or brain stem, traveling in the same nerve trunks along with the sympathetic or parasympathetic nerves. For example, 80 per cent of the nerve fibers in the vagus nerves are afferent rather than efferent. These fibers transmit afferent signals into the medulla, which in turn initiates many vagal efferent signals returning to the gastrointestinal tract to control many of its functions.

The Gastrointestinal Reflexes. The anatomical arrangement of the enteric nervous system and its connections with the sympathetic and parasympathetic systems supports three different types of gastrointestinal reflexes that are essential to gastrointestinal control. These are:

(1) *Reflexes that occur entirely within the enteric nervous system.* These include reflexes that control gastrointestinal secretion, peristalsis, mixing contractions, local inhibitory effects, and so forth.

(2) *Reflexes from the gut to the prevertebral sympathetic ganglia and then back to the gastrointestinal tract.* These reflexes transmit signals for long distances in the gastrointestinal tract, such as signals from the stomach to cause evacuation of the colon (the gastrocolic reflex), signals from the colon and small intestine to inhibit stomach motility and stomach secretion (the enterogastric reflexes), and reflexes from the colon to inhibit emptying of ileal contents into the colon (the colonoileal reflex)

(3) *Reflexes from the gut to the spinal cord or brain stem and then back to the gastrointestinal tract.* These include especially reflexes from the stomach and duodenum to the brain stem and back to the stomach to control gastric motor and secretory activity; pain reflexes that cause general inhibition throughout the gastrointestinal tract; and defecation reflexes that produce the powerful colonic, rectal, and abdominal contractions required for defecation (the defecation reflexes)

HORMONAL CONTROL OF GASTROINTESTINAL MOTILITY

In the following chapter we shall discuss in detail the extreme importance of several hormones for controlling gastrointestinal secretion Most of these same hormones also affect the motility of some parts of the gastrointestinal tract Though these effects are much less important than the secretory effects of the hormones, some of the more important of them are the following

Cholecystokinin, which is secreted mainly by the mucosa of the jejunum in response to the presence of fatty substances in the intestinal contents, has an extremely potent effect in increasing contractility of the gallbladder, thus expelling bile into the small intestine where the bile then plays important roles in emulsifying fatty substances so that they can be digested and in causing fat absorption Cholecystokinin also inhibits stomach motility moderately Therefore, at the same time that this hormone causes emptying of the gallbladder, it also slows the emptying of food from the stomach to give adequate time for digestion of the fats in the upper intestinal tract

Secretin, which is secreted by the mucosa of the duodenum in response to acidic gastric juice emptied from the stomach through the pylorus, has a mild inhibitory effect on the motility of most of the gastrointestinal tract

Gastric inhibitory peptide, which is secreted by the mucosa of the upper small intestine, mainly in response to fat but to a lesser extent in response to carbohydrate, has a mild effect in decreasing motor activity of the stomach and therefore slowing the emptying of gastric contents into the duodenum when the upper small intestine is already oversupplied with food products.

FUNCTIONAL TYPES OF MOVEMENTS IN THE GASTROINTESTINAL TRACT

Two basic types of movements occur in the gastrointestinal tract: (1) *mixing movements*, which keep the intestinal contents thoroughly mixed at all times, and (2) *propulsive movements*, which cause food to move forward along the tract at an appropriate rate for digestion and absorption.

THE MIXING MOVEMENTS

In most parts of the alimentary tract, the mixing movements are caused by either *peristaltic contractions* or *local constrictive contractions of small segments of the gut wall.* These movements are modified in different parts of the gastrointestinal tract for proper performance of the respective activities of each part, as discussed separately later in the chapter

THE PROPULSIVE MOVEMENTS—PERISTALSIS

The basic propulsive movement of the gastrointestinal tract is peristalsis, which is illustrated in Figure 63–4 A contractile ring appears around the gut and then moves forward, this is analogous to putting one's fingers around a thin distended tube, then constricting the fingers and moving forward along the tube. Obviously, any material in front of the contractile ring is moved forward.

Peristalsis is an inherent property of many syncytial smooth muscle tubes; stimulation at any point can cause a contractile ring to spread along the tube Thus, peristalsis occurs in (a) the gastrointestinal tract, (b) the bile ducts, (c) other glandular ducts throughout the body, (d) the ureters, and (e) many other smooth muscle tubes of the body.

The usual stimulus for peristalsis is *distension.* That is, if a large amount of food collects at any point in the gut, the distension stimulates the gut wall 2 to 3 cm above this point, and a contractile ring appears that initiates a peristaltic movement. Other stimuli that can initiate peristalsis include irritation of the epithelium lining the gut and extrinsic nervous signals that excite the gut

Function of the Myenteric Plexus in Peristalsis. Peristalsis occurs only weakly, if at all, in portions of the gastrointestinal tract that have congenital absence of the myenteric plexus Also, it is greatly depressed or completely blocked in the entire gut when the person is treated with atropine to paralyze the cholinergic nerve endings of the

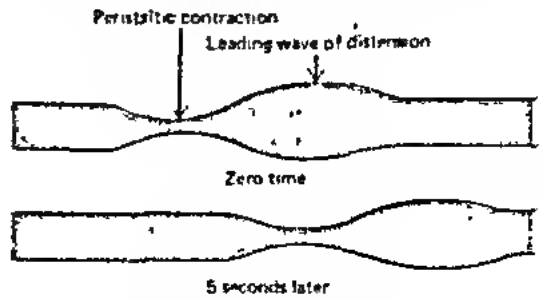


Figure 63–4. Peristalsis.

myenteric plexus. Therefore, *effectual* peristalsis requires an active myenteric plexus.

Analward Peristaltic Movements. Peristalsis, theoretically, can occur in either direction from a stimulated point, but it normally dies out rapidly in the orad direction while continuing for a considerable distance analward. The exact cause of this directional transmission of peristalsis has never been ascertained, though it probably results mainly from the fact that the myenteric plexus itself is "polarized" in the anal direction, which can be explained as follows:

The Peristaltic Reflex and the "Law of the Gut." When a segment of the intestinal tract is excited by distension and thereby initiates peristalsis, the contractile ring for causing the peristalsis begins slightly on the orad side of the distended segment; then it moves toward the distended segment, thus pushing the intestinal contents in the anal direction. At the same time, the gut sometimes relaxes several centimeters downstream toward the anus, which is called "receptive relaxation," thus allowing the food to be propelled more easily analward than in the orad direction.

This complex pattern, consisting of contraction of the gut above the point of distension and relaxation below, does not occur in the absence of the myenteric plexus. Therefore, this complex is frequently called the *myenteric reflex*, or it is also called simply the *peristaltic reflex*. And the peristaltic reflex plus the analward direction of movement of the peristalsis is called the "law of the gut."

Peristalsis in the Fasting Human Being—The Migrating Myoelectric Motor Complex

After a person or an animal eats a meal, the nature of the gastrointestinal motor functions is determined mainly by the stimulating effects of the food in the gastrointestinal tract itself. However, late at night, many hours after a meal, or at other times when the human being is fasting, a distinct pattern of activity occurs in the stomach and small intestine called the *migrating myoelectric motor complex*. The most important feature of this migrating complex is that it causes peristaltic waves to sweep slowly and rhythmically downward along the stomach and small intestine approximately once every two hours, sweeping the excess digestive secretions into the colon and therefore preventing their accumulation in the upper gastrointestinal tract.

The migrating complex begins in the body of the stomach and spreads all the way through the ileum. At any one time only about 40 cm of the intestinal tract is actively engaged in the peristaltic waves, but this 40 cm area moves slowly along the intestinal tract at a velocity of 6 to 12 cm per minute; a total period of about two hours is required in the human being for each set of migrating waves to pass the entire distance to the terminal ileum. Then approximately at the time that one migrating complex reaches the end of the ileum, a new one begins in the stomach.

The active phase of the migrating complex in any given segment of the intestine usually lasts for 10 to 15 minutes. During this time, about 50 peristaltic waves move through this portion of the intestine, that is, through this distance of about 40 cm. Thus, this represents a "sweeping" action by the waves, sweeping the intestinal contents along with the migrating complex.

INGESTION OF FOOD

The amount of food that a person ingests is determined principally by the intrinsic desire for food called *hunger*. The type of food that a person preferentially seeks is determined by *appetite*. These mechanisms in themselves are extremely important automatic regulatory systems for maintaining an adequate nutritional supply for the body, and they will be discussed in Chapter 73 in relation to nutrition of the body. The present discussion is confined to the actual mechanical aspects of food ingestion, including especially *mastication* and *swallowing*.

MASTICATION (CHEWING)

The teeth are admirably designed for chewing, the anterior teeth (incisors) providing a strong cutting action and the posterior teeth (molars) a grinding action. All the jaw muscles working together can close the teeth with a force as great as 55 pounds on the incisors and 200 pounds on the molars. When this is applied to a small object, such as a small seed between the molars, the actual force *per square inch of contact* may be several thousand pounds.

Most of the muscles of chewing are innervated by the motor branch of the 5th cranial nerve, and the chewing process is controlled by nuclei in the hindbrain. Stimulation of the reticular formation near the hindbrain centers for taste can cause continual rhythmic chewing movements. Also, stimulation of areas in the hypothalamus, amygdala, and even in the cerebral cortex near the sensory areas for taste and smell can cause chewing.

Much of the chewing process is caused by the *chewing reflex*, which may be explained as follows: The presence of a bolus of food in the mouth causes reflex inhibition of the muscles of mastication, which allows the lower jaw to drop. The drop in turn initiates a stretch reflex of the jaw muscles that leads to *rebound* contraction. This automatically raises the jaw to cause closure of the teeth, but it also compresses the bolus again against the linings of the mouth, which inhibits the jaw muscles once again, allowing the jaw to drop and rebound another time, and this is repeated again and again.

Chewing of the food is important for digestion of all foods, but it is especially important for most fruits and raw vegetables, because these have undigestible cellulose membranes aro[illegible] trient portions which must be broken [illegible]

food can be utilized. Chewing aids in the digestion of food for the following simple reason: Since the *digestive enzymes act only on the surfaces of food particles*, the rate of digestion is highly dependent on the total surface area exposed to the intestinal secretions. Also, grinding the food to a very fine particulate consistency prevents excoriation of the gastrointestinal tract and increases the ease with which food is emptied from the stomach into the small intestine and thence into all succeeding segments of the gut.

SWALLOWING (DEGLUTITION)

Swallowing is a complicated mechanism, principally because the pharynx most of the time subserves several other functions besides swallowing and is converted for only a few seconds at a time into a tract for propulsion of food. Especially is it important that respiration not be seriously compromised during swallowing.

In general, swallowing can be divided into (1) the *voluntary stage*, which initiates the swallowing process, (2) the *pharyngeal stage*, which is involuntary and constitutes the passage of food through the pharynx into the esophagus, and (3) the *esophageal stage*, another involuntory phase which promotes passage of food from the pharynx to the stomach.

Voluntary Stage of Swallowing. When the food is ready for swallowing, it is "voluntarily" squeezed or rolled posteriorly in the mouth by pressure of the tongue upward and backward against the palate, as shown in Figure 63–5. Thus, the tongue forces the bolus of food into the pharynx. From here on, the process of swallowing becomes entirely, or almost entirely, automatic and ordinarily cannot be stopped.

Pharyngeal Stage of Swallowing. When the bolus of food is pushed backward in the mouth, it stimulates *swallowing receptor areas* all around the opening of the pharynx, especially on the tonsillar pillars, and impulses from these pass to the brain stem to initiate a series of automatic pharyngeal muscular contractions as follows:

1. The soft palate is pulled upward to close the posterior nares, in this way preventing reflux of food into the nasal cavities.

2. The palatopharyngeal folds on either side of the pharynx are pulled medialward to approximate each other. In this way these folds form a sagittal slit through which the food must pass into the posterior pharynx. This slit performs a selective action, allowing food that has been masticated properly to pass with ease while impeding the passage of large objects. Since this stoge of swallowing lasts less than 1 second, any large object is usually impeded too much to pass through the pharynx into the esophagus.

3. The vocal cords of the larynx are strongly [illegible] and the hyoid bone and larynx are pulled upward and anteriorly by the neck muscles, causing the epiglottis to swing backward over the superior opening of the larynx. Both these effects prevent passage of food into the trachea. Especially important is the approximation of the vocal cords, but the epiglottis helps to prevent food from ever getting as far as the vocal cords. Destruction of the vocal cords or of the muscles that approximate them can cause strangulation. On the other hand, removal of the epiglottis usually does not cause serious debility in swallowing.

4. The upward movement of the larynx also stretches the opening of the esophagus. At the same time, the upper 3 to 4 centimeters of the esophagus, an area called the *upper esophageal sphincter*, the *pharyngoesophageal sphincter*, or the *cricopharyngeal muscle*, relaxes, thus allowing food to move easily and freely from the posterior pharynx into the upper esophagus. This sphincter, between swallows, remains tonically and strongly contracted, thereby preventing air from going into the esophagus during respiration. The upward movement of the larynx also lifts the glottis out of the main stream of food flow so that the food usually passes on either side of the epiglottis rather than over its surfsce; this adds still another protection against passage of food into the traches.

5. At the same time that the larynx is raised and the pharyngoesophageal sphincter is relaxed, the superior constrictor muscle of the phsrynx contracts, giving rise to a rapid peristaltic wave passing downwsrd over the middle and inferior pharyngeal muscles and into the esophagus, which also propels the food into the esophagus.

To summarize the mechanics of the phsryngeal stage of swsllowing—the trachea is closed, the esophagus is opened, and a fast peristaltic wave originating in the pharynx then forces the bolus

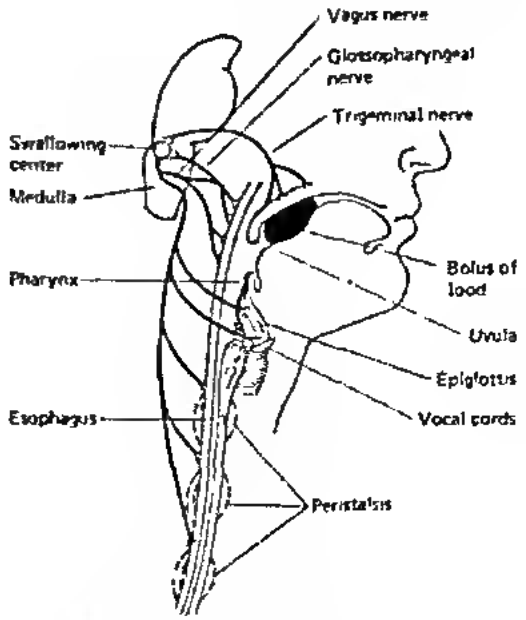


Figure 63–5. The swallowing mechanism.

of food into the upper esophagus, the entire process occurring in 1 to 2 seconds.

Nervous Control of the Pharyngeal Stage of Swallowing. The most sensitive tactile areas of the pharynx for initiation of the pharyngeal stage of swallowing lie in a ring around the pharyngeal opening, with greatest sensitivity in the tonsillar pillars. Impulses are transmitted from these areas through the sensory portions of the trigeminal and glossopharyngeal nerves into a region of the medulla oblongata closely associated with the *tractus solitarius* which receives essentially all sensory impulses from the mouth.

The successive stages of the swallowing process are then automatically controlled in orderly sequence by neuronal areas distributed throughout the reticular substance of the medulla and lower portion of the pons. The sequence of the swallowing reflex is the same from one swallow to the next, and the timing of the entire cycle also remains constant from one swallow to the next. The areas in the medulla and lower pons that control swallowing are collectively called the *deglutition* or *swallowing center.*

The motor impulses from the swallowing center to the pharynx and upper esophagus that cause swallowing are transmitted by the 5th, 9th, 10th, and 12th cranial nerves and even a few of the superior cervical nerves.

In summary, the pharyngeal stage of swallowing is principally a reflex act. It is almost never initiated by direct stimuli to the swallowing center from higher regions of the central nervous system. Instead, it is almost always initiated by voluntary movement of food into the back of the mouth, which, in turn, elicits the swallowing reflex.

Effect of the Pharyngeal Stage of Swallowing on Respiration. The entire pharyngeal stage of swallowing occurs in less than 1 to 2 seconds, thereby interrupting respiration for only a fraction of a usual respiratory cycle. The swallowing center specifically inhibits the respiratory center of the medulla during this time, halting respiration at any point in its cycle to allow swallowing to proceed. Yet, even while a person is talking, swallowing interrupts respiration for such a short time that it is hardly noticeable.

Esophageal Stage of Swallowing. The esophagus functions primarily to conduct food from the pharynx to the stomach, and its movements are organized specifically for this function.

Normally the esophagus exhibits two types of peristaltic movements—*primary peristalsis* and *secondary peristalsis.* Primary peristalsis is simply a continuation of the peristaltic wave that begins in the pharynx and spreads into the esophagus during the pharyngeal stage of swallowing. This wave passes all the way from the pharynx to the stomach in approximately 8 to 10 seconds. However, food swallowed by a person who is in the upright position is usually transmitted to the lower end of the esophagus even more rapidly than the peristaltic wave itself, in about 5 to 8 seconds, because of the additional effect of gravity pulling the food downward. If the primary peristaltic wave fails to move all the food that has entered the esophagus into the stomach, secondary peristaltic waves, generated by the enteric nervous system of the esophagus, result from distension of the esophagus by the retained food. These waves are essentially the same as the primary peristaltic waves, except that they originate in the esophagus itself rather than in the pharynx. Secondary peristaltic waves continue to be initiated until all the food has emptied into the stomach.

The peristaltic waves of the esophagus are initiated by vagal reflexes that are part of the overall swallowing mechanism. These reflexes are transmitted through *vagal afferent fibers* from the esophagus to the medulla and then back again to the esophagus through *vagal afferent fibers.*

The musculature of the pharynx and the upper quarter of the esophagus is striated muscle, and, therefore, the peristaltic waves in these regions are controlled only by skeletal nerve impulses in the glossopharyngeal and vagus nerves. In the lower two thirds of the esophagus, the musculature is smooth, but this portion of the esophagus is also strongly controlled by the vagus nerves acting through their connections with the enteric nervous system. However, when the vagus nerves to the esophagus are sectioned, the myenteric nerve plexus of the esophagus becomes excitable enough after several days to cause secondary peristaltic waves even without support from the vagal reflexes. Therefore, following paralysis of the swallowing reflex, food forced into the upper esophagus and then pulled by gravity to the lower esophagus still passes readily into the stomach.

Receptive Relaxation of the Stomach. As the esophageal peristaltic wave passes toward the stomach, a wave of relaxation, transmitted through myenteric inhibitory neurons, precedes the constriction. Furthermore, the entire stomach and, to a lesser extent, even the duodenum become relaxed as this wave reaches the lower end of the esophagus. Especially important, also, is relaxation of the gastroesophageal sphincter at the juncture between the esophagus and the stomach. In other words, the constrictor and the stomach are prepared ahead of time to receive food being propelled down the esophagus during the swallowing act.

FUNCTION OF THE LOWER ESOPHAGEAL SPHINCTER (GASTROESOPHAGEAL SPHINCTER)

At the lower end of the esophagus, extending from about 2 to 5 cm above its juncture with the stomach, the circular muscle functions as a so-called *lower esophageal sphincter* or *gastroesopha-*

geal sphincter. Anatomically this sphincter is no different from the remainder of the esophagus. However, physiologically, it normally remains tonically constricted, in contrast to the mid- and upper portions of the esophagus which normally remain completely relaxed. However, when a peristaltic wave of swallowing passes down the esophagus, "receptive relaxation" relaxes the lower esophageal sphincter ahead of the peristaltic wave, and allows easy propulsion of the swallowed food into the stomach. Rarely, the sphincter does not relax satisfactorily, resulting in a condition called *achalasia*, which will be discussed in Chapter 66.

A principal function of the lower esophageal sphincter is to prevent reflux of stomach contents into the esophagus. The stomach contents are highly acidic and contain many proteolytic enzymes The esophageal mucosa, except in the lower eighth of the esophagus, is not capable of resisting for long the digestive action of gastric secretions. Fortunately, the tonic constriction of the lower esophageal sphincter prevents significant reflux of stomach contents into the esophagus except under abnormal conditions Indeed, increased intragastric pressure, except during retching or vomiting, causes a vagal reflex that further constricts the sphincter to add extra insurance against reflux

Prevention of Reflux by Flutter-Valve Closure of the Distal End of the Esophagus. Another factor that prevents reflux is a valvelike mechanism of that portion of the esophagus that lies immediately beneath the diaphragm. Greatly increased intra-abdominal pressure caves the esophagus inward at this point at the same time that the abdominal pressure also increases the intragastric pressure. This flutter-valve closure of the lower esophagus therefore prevents the high pressure in the stomach from forcing stomach contents into the esophagus. Otherwise, every time we should walk, cough, or breath hard, we might expel acid into the esophagus.

MOTOR FUNCTIONS OF THE STOMACH

The motor functions of the stomach are threefold: (1) storage of large quantities of food until it can be accommodated in the lower portion of the gastrointestinal tract, (2) mixing of this food with gastric secretions until it forms a semifluid mixture called *chyme*, and (3) slow emptying of the food from the stomach into the small intestine at a rate suitable for proper digestion and absorption by the small intestine.

Figure 63–6 illustrates the basic anatomy of the stomach. Physiologically, the stomach can be divided into two major parts: (1) the *corpus*, or *body*, and (2) the *antrum*. The *fundus*, located at the upper end of the body of the stomach, is considered some anatomists to be a separate entity from the body, but physiologically the fundus functions mainly as part of the body.

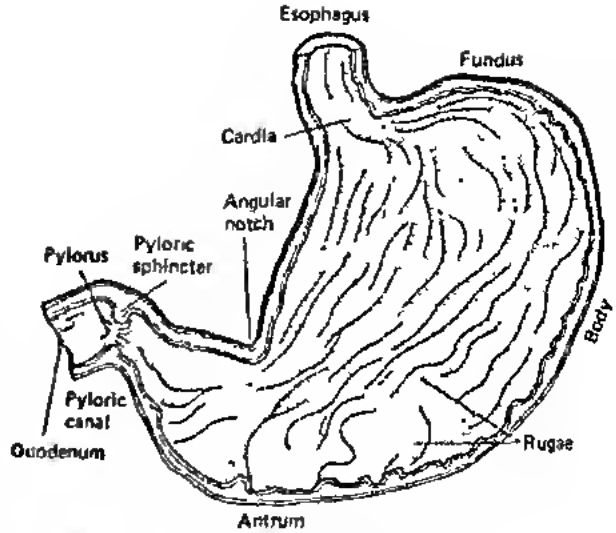


Figure 63–6. Physiologic anatomy of the stomach

STORAGE FUNCTION OF THE STOMACH

As food enters the stomach, it forms concentric circles in the body and fundus of the stomach, the newest food lying closest to the esophageal opening and the oldest food lying nearest the wall of the stomach. Normally, when food enters the stomach, a vagal reflex greatly reduces the tone in the muscular wall of tha body of the stomach so that the wall can bulge progressively outward, accommodating greater and greater quantities of food up to a limit of about 1 liter. The pressure in the stomach remains low until this limit is approached.

MIXING AND PROPULSION OF FOOD IN THE STOMACH—THE BASIC ELECTRICAL RHYTHM OF THE STOMACH

The digestive juices of the stomach are secreted by the *gastric glands*, which cover almost the entire outer wall of the body of the stomach. These secretions come immediately into contact with that portion of the stored food lying against the murosal surface of the stomach; when the stomach is filled, weak peristaltic *constrictor waves*, also called *mixing waves*, move toward the antrum along the stomach wall approximately once every 20 seconds. These waves are controlled by the *basic electrical rhythm* (called *BER*) consisting of electrical "slow waves" that occur spontaneously in the stomach wall. As the waves move down the stomach, they not only cause the secretions to mix with the stored food but also provide weak propulsion to move these mixed contents into the antrum. When the stomach is full, these mixing waves usually begin near the midpoint of the stomach, but, as the stomach empties, the waves become stronger and

also originate farther back up the stomach wall, thus propelling the last vestiges of stored food into the stomach antrum.

As the constrictor waves progress from the body of the stomach into the antrum, they become more intense, some becoming extremely intense and providing powerful *peristaltic constrictor rings* that force the antral contents under high pressure toward the pylorus. These constrictor rings also play an exceedingly important role in mixing the stomach contents in the following way: Each time a peristaltic wave passes over the antrum toward the pylorus, it digs deeply into the contents of the antrum. Yet the opening of the pylorus is small enough that only a few milliliters of antral contents are expelled into the duodenum with each peristaltic wave. Also, as each peristaltic wave approaches the pylorus, the pyloric muscle itself contracts, which further impedes emptying through the pylorus. Therefore, most of the antral contents are squirted backward through the peristaltic ring toward the body of the stomach. Thus, the moving peristaltic constrictive ring, combined with this squirting action, called "retropulsion," is an exceedingly important mixing mechanism of the stomach.

Chyme. After the food has become mixed with the stomach secretions, the resulting mixture that passes on down the gut is called *chyme*. The degree of fluidity of chyme depends on the relative amounts of food and stomach secretions and on the degree of digestion that has occurred. The appearance of chyme is that of e murky, milky semifluid or paste.

Hunger Contractions. Besides the peristaltic contractions that occur when food is present in the stomach, another type of intense contractions, called *hunger contractions*, often occurs when the stomach has been empty for a long time. These are usually rhythmic peristaltic contractions probably representing exacerbated peristaltic mixing waves in the *body* of the stomach. However, when they become extremely strong, they often fuse together to cause a continuing tetanic contraction lasting for as long as two to three minutes.

Hunger contractions are usually most intense in young healthy persons with high degrees of gastrointestinal tonus, and they are also greatly increased by a low level of blood sugar.

When hunger contractions occur in the stomach, the person sometimes experiences a sensation of pain in the pit of the stomach, called *hunger pangs*. Hunger pangs usually do not begin until 12 to 24 hours after the last ingestion of food; in starvation they reach their greatest intensity in three to four days, and then gradually weaken in succeeding days.

Hunger contractions are often associated with a feeling of hunger and therefore are perhaps an important means by which the alimentary tract intensifies the animal drive to acquire food when a person is in a state of incipient starvation.

EMPTYING OF THE STOMACH

Basically, stomach emptying is opposed by resistance of the pylorus to the passage of food, and it is promoted by peristaltic waves in the antrum of the stomach.

Role of the Pylorus in Stomach Emptying. The pylorus normally remains almost, but not completely, closed because of tonic contraction of the pyloric muscle. The closing force is weak enough that water and other fluids empty from the stomach with ease. On the other hand, it is great enough to prevent movement of semi-solid chyme into the duodenum except when a strong antral peristaltic wave forces the chyme through. However, the degree of constriction of the pyloric sphincter can increase or decrease under the influence of signals both from the stomach and from the duodenum, as we shall discuss subsequently. In this way, the pylorus plays an important role in the control of stomach emptying.

Role of Antral Peristalsis in Stomach Emptying—The Pyloric Pump. The intensity of antral peristalsis changes markedly under different conditions, especially in response to signals both from the stomach and from the duodenum. Most of the time the antral peristaltic contractions are weak and function mainly to cause mixing of the food and gastric secretions, thus increasing the fluidity of the chyme. However, about 20 per cent of the time while food is in the stomach, these peristaltic contractions become very intense at the incisura angularis of the stomach and spread through the antrum no longer as weak mixing waves, but instead as strong peristaltic, ringlike constrictions. As the stomach becomes progressively more and more empty, these constrictions begin farther and farther up the body of the stomach, gradually pinching off the lowermost portions of the stored food and adding this food to the chyme in the antrum. These intense peristaltic waves often create as much as 50 to 70 centimeters of water pressure, which is about six times as powerful as the usual mixing type of peristaltic waves. Thus, the intensity of this antral peristelsis is the second principal factor determining the rate of stomach emptying.

When pyloric tone is normal, each strong antral peristaltic wave forces several milliliters of chyme into the duodenum. Thus, the peristaltic waves provide a pumping action that is frequently called the "pyloric pump."

REGULATION OF STOMACH EMPTYING

The rate at which the stomach empties is regulated by signals both from the stomach and from the duodenum. The stomach signals are mainly twofold: (1) nervous signals caused by distension of the stomach by food, and (2) the hormone *gastrin* released from the antral mucosa in response to the presence of certain types of food in the stomach. Both these signals increase pyloric pumping force and at the same time inhibit the pylorus, thus promoting stomach emptying.

On the other hand, signals from the duodenum

depress the pyloric pump and usually increase pyloric tone at the same time. In general, when an excess volume of chyme or excesses of certain types of chyme enter the duodenum, strong *negative* feedback signals, both *nervous* and *hormonal*, depress the pyloric pump and enhance pyloric sphincter tone. Obviously, these feedback signals allow chyme to enter the duodenum only as rapidly as it can be processed by the small intestine.

Gastric Factors That Promote Emptying

Effect of Gastric Food Volume on Rate of Emptying. It is very easy to see how increased food volume in the stomach could promote increased emptying from the stomach. However, this increased emptying does not occur for the reasons that one would expect. It is not increased pressure in the stomach that causes the increased emptying because, in the usual normal range of volume, the increase in volume does not increase the pressure significantly On the other hand, stretch of the stomach wall does elicit vagal and local myenteric reflexes in the wall that increase the activity of the pyloric pump and at the same time inhibit the pylorus In general, the rate of food emptying from the stomach is approximately proportional to the *square root* of the volume of food remaining in the stomach at any given time

Effect of the Hormone Gastrin on Stomach Emptying. In the following chapter we shall see that stretch, as well as the presence of certain types of foods in the stomach—particularly meat—elicits release of a hormone called *gastrin* from the antral mucosa, and this has potent effects on causing secretion of highly acidic gastric juice by the stomach fundic glands However, gastrin also has moderate stimulatory effects on motor functions of the stomach as well Most important, it enhances the activity of the pyloric pump while at the same time relaxing the pylorus Thus, it is an important influence for promoting stomach emptying It also has a slight constrictor effect on the gastroesophageal sphincter at the lower end of the esophagus for preventing reflux of gastric contents into the esophagus during the enhanced gastric activity

Duodenal Factors That Inhibit Emptying

The Inhibitory Effect of the Enterogastric Reflex from the Duodenum on Pyloric Activity. Reflex nervous signals are transmitted from the duodenum back to the stomach most of the time when the stomach is emptying food into the duodenum These signals play an especially important role in controlling both the pyloric pump and the pylorus and, therefore, also in determining the rate of emptying of the stomach. The nervous reflexes are mediated partly through the extrinsic [illegible]es, some going to the prevertebral sympathetic [illegible] and then returning through inhibitory sympathetic nerve fibers to the stomach, and others transmitted in the vagus nerves to the brain stem and then back through the same vagi to the stomach However, local signals are probably also transmitted from the upper portion of the duodenum to the pylorus directly through the enteric nervous system within the gut wall itself.

The types of factors that are continually monitored in the duodenum and that can elicit the enterogastric reflex include:

1. The degree of distension of the duodenum
2. The presence of any degree of irritation of the duodenal mucosa
3. The degree of acidity of the duodenal chyme
4. The degree of osmolality of the chyme.
5. The presence of certain breakdown products in the chyme, especially breakdown products of proteins and perhaps to a lesser extent of fats.

The enterogastric reflex is especially sensitive to the presence of irritants and acids in the duodenal chyme For instance, whenever the pH of the chyme in the duodenum falls below approximately 3 5 to 4, this reflex is immediately elicited, which inhibits the pyloric pump and increases pyloric constriction, thus reducing or even blocking further release of acidic stomach contents into the duodenum until the duodenal chyme can be neutralized by pancreatic and other secretions.

Breakdown products of protein digestion will also elicit this reflex, by slowing the rate of stomach emptying, sufficient time is insured for adequate protein digestion in the upper portion of the small intestine

Finally, either hypo- or hypertonic fluids (especially hypertonic) will elicit the enterogastric reflex This effect prevents too rapid flow of nonisotonic fluids into the small intestine, thereby preventing rapid changes in electrolyte balance of the body fluids during absorption of the intestinal contents

Hormonal Feedback from the Duodenum in Inhibiting Gastric Emptying—Role of Fats. Not only do nervous reflexes from the duodenum to the stomach inhibit stomach emptying, but hormones released from the upper intestine do so as well The stimulus for producing the hormones is mainly fats entering the duodenum, though other types of foods can stimulate the hormones to a lesser degree On entering the duodenum, the fats extract several different hormones from the duodenal and jejunal epithelium, and these in turn are carried by way of the blood to the stomach where they (a) inhibit the activity of the pyloric pump and at the same time (b) increase slightly the strength of contraction of the pyloric sphincter These effects are important because fats are much slower to be digested than are most other foods; this inhibitory feedback mechanism allows the necessary time before the fats pass deeper into the intestinal tract where they are to be absorbed.

Unfortunately, the precise hormones that cause

the hormonal feedback inhibition of the stomach are not fully clear. In the past, this mixture of hormones has been called *enterogastrone*, but such a single hormone has never been identified as a specific entity. On the other hand, several different hormones released by the mucosa of the upper small intestine are known to inhibit stomach emptying. One of these is *cholecystokinin*, which is released from the mucosa of the jejunum in response to fatty substances in the chyme. This hormone acts as a competitive inhibitor to block the increased stomach motility caused by gastrin. Another ie the hormone *secretin*, which is released mainly from the duodenal mucosa in response to gastric acid released from the stomach through the pylorus. This hormone has the general but only weak effect of decreasing gastrointestinal motility. Finally, a hormone called *gastric inhibitory peptide*, which is released from the upper small intestine in response mainly to fat in the chyme but to carbohydrates as well, is known also to inhibit gastric motility under some conditions (However, its effect at physiological concentrations is probably mainly to stimulate the secretion of insulin by the pancreas) All these hormones will bo discussed at greater length elsewhere in this text, especially in tha following chapter where both cholecystokinin and secretin will be discussed in detoil.

In summary, several different hormones are known that could serve os hormonal mechanisms for inhibiting gastric emptying when excess quantities of chyme, especially acidic or fatty chyme, enter the duodenum from the stomach.

Summary of the Control of Stomach Emptying

Emptying of the stomach is controlled to a moderate degree by stomach factors, such as the degree of filling in the stomach and the excitatory effect of gastrin on antral peristalsis. However, probably the more important control of stomach emptying resides in feedback signals from the duodenum, including both the enterogastric feedback reflex and hormonal feedback. These two feedback inhibitory signals work together to slow the rate of emptying when (a) too much chyme is already in the small intestine or (b) the chyme is excessively acid, contains too much protein or fat, is hypotonic or hypertonic, or is irritating. In this way the rate of stomach emptying is limited to thet amount of chyme that the small intestine can process.

MOVEMENTS OF THE SMALL INTESTINE

The movements of the small intestine, as elsewhere in the gastrointestinal tract, can be divided into the *mixing contractions* and the *propulsive contractions*. However, to a great extent this separation is artificial because essentielly all movements of the small intestine cause at least some degree of both mixing and propulsion. Yet, the usual classification of these processes is the following:

MIXING CONTRACTIONS (SEGMENTATION CONTRACTIONS)

When a portion of the small intestine becomes distended with chyme, the stretch of the intestinal wall elicits localized concentric contractions spaced at intervals along the intestine. The longitudinal length of each one of the contractions is only about 1 cm so that each set of contractions causes "segmentation" of the small intestine, as illustrated in Figure 63–7, dividing the intestine into spaced segments that have the appearance of a chain of sausages As one set of segmentation contractions relaxes a new set begins, but the contractions this time occur at new points between the previous contractions. These segmentation contrections "chop" the chyme as often es 8 to 12 times a minute, in this way promoting progreseive mixing of the solid food particles with the secretions of the small intestine.

The maximum frequency of the segmentetion contractions in the small intestine is determined by the frequency of the *slow waves* in the intestinal wall, which is the basic electrical rhythm (BER) as explained earlier Since this frequency is about 12 per minute in the duodenum, the maximum frequency of the segmentation contractions in tha duodenum is also about 12 per minute. However, in the ileum, the maximum frequency is usually 8 to 9 contractions per minute

The segmentation contractions are exceedingly weak when the excitatory activity of the enteric nervous system is blocked by atropine. Therefore, even though it is the slow waves in the smooth muscle itself that control the segmentation contractions, these contractions are not really effective without background excitation by the enteric nervous system, especially by the myenteric plexis.

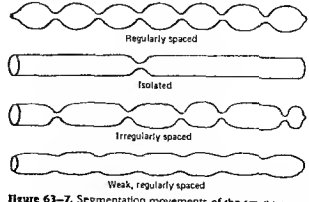


Figure 63–7. Segmentation movements of the small intestine.

PROPULSIVE MOVEMENTS

Peristalsis in the Small Intestine. Chyme is propelled through the small intestine by by *peristaltic waves*. These can occur in any part of the small intestine, and they move analward at a velocity of 0 5 to 2 cm per second, much faster in the proximal intestine and much slower in the terminal intestine. However, they are normally very weak and usually die out after traveling less than 10 cm, so that movement of the chyme is also very poor, so poor in fact that the *net* movement of the chyme along the small intestine averages only 1 cm per minute. This means that normally 3 to 5 hours are required for passage of chyme from the pylorus to the ileocecal valve

Peristaltic activity of the small intestine is greatly increased after a meal This is caused partly by the beginning entry of chyme into the duodenum but also by the so-called *gastroenteric reflex* that is initiated by distension of the stomach and conducted principally through the myenteric plexus from the stomach down along the wall of the small intestine This reflex increases the overall degree of excitability of the small intestine, including both increased motility and secretion.

The function of the peristaltic waves in the small intestine is not only to cause progression of the chyme toward the ileocecal valve but also to spread out the chyme along the intestinal mucosa As the chyme enters the intestine from the stomach and causes initial distension of the proximal intestine, the elicited peristaltic waves begin immediately to spread the chyme along the intestine, and this process intensifies as additional chyme enters the intestine On reaching the ileocecal valve the chyme is sometimes blocked for several hours until the person eats another meal, when a new *gastroenteric* (also called *gastroileal*) reflex intensifies the peristalsis in the ileum and forces the remaining chyme through the ileocecal valve into the cecum

The Propulsive Effect of the Segmentation Movements. The segmentation movements, though they last for only a few seconds, also travel in the analward direction and help propel the food down the intestine Therefore, the difference between the segmentation and the peristaltic movements is not as great as might be implied by their separation into these two classifications.

The Peristaltic Rush. Though peristalsis in the small intestine is normally very weak, intense irritation of the intestinal mucosa, as occurs in some severe cases of infectious diarrhea, can cause both very powerful and rapid peristalsis called the *peristaltic rush*. This is initiated mainly by extrinsic nervous reflexes to the brain stem and back again to the gut. The powerful peristaltic contractions then travel long distances in the small intes- within minutes, sweeping the contents of the into the colon and thereby relieving the small intestine of either irritative chyme or excessive distension.

Movements Caused by the Muscularis Mucosae and Muscle Fibers of the Villi

The muscularis mucosae, which is stimulated by local nervous reflexes in the submucosal plexus, can cause short or long folds to appear in the intestinal mucosa and also cause the folds to move to progressively newer areas of mucosa. Also, individual fibers from this muscle extend into the intestinal villi and cause them to contract intermittently The mucosal folds increase the surface area exposed to the chyme, thereby increasing the rate of absorption. The contractions of the villi—shortening, elongating, and shortening again—"milk" the villi so that lymph flows freely from the central lacteals into the lymphatic system. Both these types of contraction also agitate the fluids surrounding the villi so that progressively new areas of fluid become exposed to absorption.

These mucosal and villous contractions are initiated by local nervous reflexes that occur in response to chyme in the small intestine.

FUNCTION OF THE ILEOCECAL VALVE

A principal function of the ileocecal valve is to prevent backflow of fecal contents from the colon into the small intestine As illustrated in Figure 63–8, the lips of the ileocecal valve protrude into the lumen of the cecum and therefore are forcefully closed when the cecum fills Usually the valve can resist reverse pressure of as much as 50 to 60 cm water

The wall of the ileum for several centimeters immediately preceding the ileocecal valve has a thickened muscular coat called the *ileocecal*

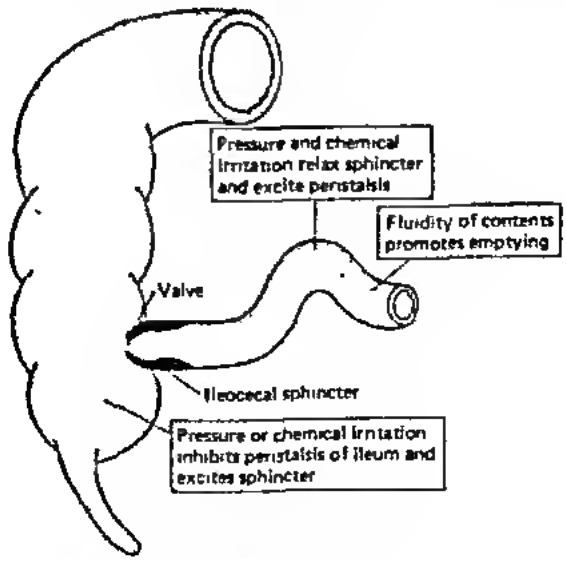


Figure 63–8. Emptying at the ileocecal valve

sphincter. This normally remains mildly constricted and slows the emptying of ileal contents into the cecum except immediately following a meal when a gastroileal reflex (described earlier) intensifies the peristalsis in the ileum. Also, the hormone *gastrin*, which is liberated from the stomach mucosa in response to food in the stomach, has a relaxant effect on the ileocecal sphincter, thus allowing increased emptying. Even so, only about 1500 ml of chyme empty into the cecum each day. The resistance to emptying at the ileocecal valve prolongs the stay of chyme in the ileum and, therefore, facilitates absorption.

Feedback Control of the Ileocecal Sphincter. The degree of contraction of the ileocecal sphincter, as well as the intensity of peristalsis in the terminal ileum, is also controlled strongly by reflexes from the cecum. Whenever the cecum is distended, the degree of contraction of the ileocecal sphincter is intensified while ileal peristalsis is inhibited, which greatly delays emptying of additional chyme from the ileum. Also, any irritant in the cecum delays emptying. For instance, when a person has an inflamed appendix, the irritation of this vestigial remnant of the cecum can cause such intense spasm of the ileocecal sphincter and paralysis of the ileum that this completely blocks emptying of the ileum. These reflexes from the cecum to the ileocecal sphincter and ileum are mediated both by way of the myenteric plexus and through extrinsic nerves, especially reflexes by way of the prevertebral sympathetic ganglia.

MOVEMENTS OF THE COLON

The functions of the colon are (1) absorption of water and electrolytes from the chyme and (2) storage of fecal matter until it can be expelled. The proximal half of the colon, illustrated in Figure 63–9, is concerned principally with absorption, and the distal half with storage. Since intense movements are not required for these functions, the movements of the colon are normally sluggish. Yet in a sluggish manner, the movements still have characteristics similar to those of the small intestine and can be divided once again into mixing movements and propulsive movements.

Mixing Movements—Haustrations. In the same manner that segmentation movements occur in the small intestine, large circular constrictions also occur in the large intestine. At each of these constriction points, about 2.5 cm of the circular muscle contracts, sometimes constricting the lumen of the colon to almost complete occlusion. At the same time, the longitudinal muscle of the colon, which is aggregated into three longitudinal strips called the *teniae coli*, contract. These combined contractions of the circular and longitudinal smooth muscle cause the unstimulated portion of the large intestine to bulge outward into baglike sacs called *haustrations*. The haustral contractions, once initiated, usually reach peak intensity in about 30 seconds and then disappear during the next 60 seconds. They also at times move slowly analward during their period of contraction, especially in the cecum and ascending colon. After another few minutes, new haustral contractions occur in other areas nearby. Therefore, the fecal material in the large intestine is slowly *dug into and rolled over* in much the same manner that one spades the earth. In this way, all the fecal material is gradually exposed to the surface of the large intestine, and fluid is progressively absorbed until only 80 to 150 ml of the 1500 ml daily load of chyme is lost in the feces.

Propulsive Movements—"Mass Movements." Peristaltic waves of the type seen in the small intestine only rarely occur in most parts of the colon. Instead, most propulsion occurs by (1) the slow analward movement of the *haustral contractions* just discussed and (2) *mass movements*.

Most of the propulsion in the cecum and ascending colon results from the slow but persistent haustral contractions, requiring as many as 8 to 15 hours to move the chyme only from the ileocecal valve to the transverse colon, while the chyme itself becomes fecal in quality and also becomes a semisolid slush instead of a semifluid.

From the transverse colon to the sigmoid, mass movements mainly take over the propulsive role. These movements usually occur only a few times each day, most abundantly for about 15 minutes during the first hour after eating breakfast.

A mass movement is a modified type of peristalsis characterized by the following sequence of events: First, a constrictive ring occurs at a distended or irritated point in the colon, usually in the transverse colon, and then rapidly thereafter the 20 or more centimeters of colon *distal* to the constriction contract almost as a unit, forcing the fecal material in this segment *en masse* down the colon. During this process, the haustrations disappear completely. The initiation of contraction is

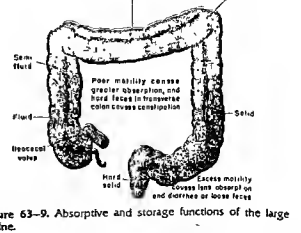


Figure 63–9. Absorptive and storage functions of the large intestine.

complete in about 30 seconds, and relaxation then occurs during the next 2 to 3 minutes before another mass movement occurs, this time perhaps farther along the colon. But the whole series of mass movements will usually persist for only 10 minutes to half an hour, and they will then return perhaps a half day or even a day later

Mass movements can occur in any part of the colon, though most often they occur in the transverse or descending colon. When they have forced a mass of feces into the rectum, the desire for defecation is felt.

Initiation of Mass Movements by the Gastrocolic and Duodenocolic Reflexes. The appearance of mass movements after meals is facilitated by *gastrocolic* and *duodenocolic reflexes*. These reflexes result from distension of the stomach and duodenum They can take place only weakly when the extrinsic nerves are removed; therefore, it is probable that weak reflexes are basically transmitted through the myenteric plexus, though reflexes conducted through the extrinsic nerves of the autonomic nervous system probably provide most of the intensity of the gastrocolic and duodenocolic reflexes

Irritation in the colon can also initiate intense mass movements For instance, a person who has an ulcerated condition of the colon *(ulcerative colitis)* frequently has mass movements that persist almost all the time

Mass movements can also be initiated by intense stimulation of the parasympathetic nervous system or simply by overdistension of a segment of the colon

DEFECATION

Most of the time, the rectum is empty of feces This results partly from the fact that a weak functional sphincter exists approximately 20 cm from the anus at the juncture between the sigmoid and the rectum There is also a sharp angulation here that contributes additional resistance to filling of the rectum However, when a mass movement forces feces into the rectum the desire for defecation is normally initiated, including reflex contraction of the rectum and relaxation of the anal sphincters.

Continual dribble of fecal matter through the anus is prevented by tonic constriction of (1) the *internal anal sphincter*, a circular mass of smooth muscle that lies immediately inside the anus, and (2) the *external anal sphincter*, composed of striated voluntary muscle that both surrounds the internal sphincter and also extends distal to it; the external sphincter is controlled by nerve fibers in the pudendal nerve, which is part of the somatic nervous system and therefore is under voluntary, conscious control.

The Defecation Reflexes. Ordinarily, defecation is initiated by *defecation reflexes*, which can be described as follows: When the feces enter the rectum, distension of the rectal wall initiates afferent signals that spread through the *myenteric plexus* to initiate peristaltic waves in the descending colon, sigmoid, and rectum, forcing feces toward the anus. As the peristaltic wave approaches the anus, the internal anal sphincter is inhibited by inhibitory nerves, and if the external anal sphincter is relaxed, defecation will occur. This overall effect is called the *intrinsic defecation reflex* of the colon itself

However, the intrinsic defecation reflex itself is usually weak, and to be effective in causing defecation it must be fortified by another type of defecation reflex, a *parasympathetic defecation reflex* that involves the sacral segments of the spinal cord, as illustrated in Figure 63–10 When the afferent fibers in the rectum are stimulated, signals are transmitted into the spinal cord and thence, reflexly, back to the descending colon, sigmoid, rectum, and anus by way of parasympathetic nerve fibers in the *pelvic nerves*. These parasympathetic signals greatly intensify the peristaltic waves as well as relaxing the internal anal sphincter and convert the intrinsic defecation reflex from an ineffectual weak movement into a powerful process of defecation that is sometimes effective in emptying the large bowel in one movement all the way from the splenic flexure to the anus Also, the afferent signals entering the spinal cord initiate other effects, such as taking a deep breath, closure of the glottis, and contraction of the abdominal muscles to force the fecal contents of the colon downward, at the same time causing the pelvic floor to extend downward to pull outward on the anus to evaginate the feces

However, despite the defecation reflexes other effects are also necessary before actual defecation occurs, because relaxation of the internal sphincter and forward movement of feces toward the anus

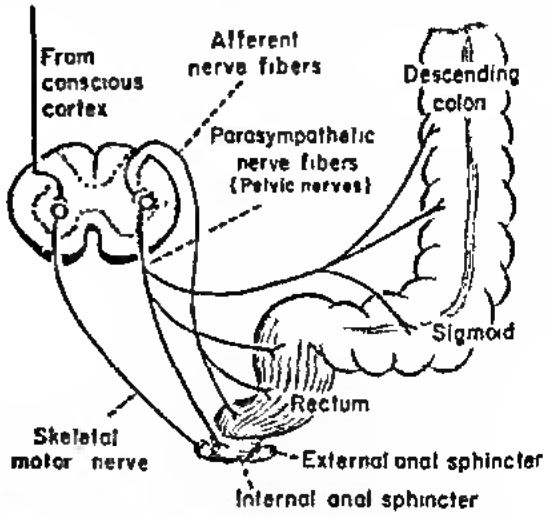


Figure 63–10. The afferent and efferent pathways of the parasympathetic mechanism for enhancing the defecation reflex.

normally initiates an instantaneous contraction of the external sphincter, which still temporarily prevents defecation. Except in babies and mentally inept persons, the conscious mind then takes over voluntary control of the external sphincter and either inhibits it to allow defecation to occur or further contracts it if the moment is not socially acceptable for defecation. If the external sphincter is kept contracted, the defecation reflexes die out after a few minutes, and they remain quiescent for several hours or until additional amounts of feces enter the rectum.

When it becomes convenient for the person to defecate, the defecation reflexes can sometimes be excited by taking a deep breath to move the diaphragm downward and then contracting the abdominal muscles to increase the pressure in abdomen, thus forcing fecal contents into the rectum to elicit new reflexes. Unfortunately, reflexes initiated in this way are never as effective as those that arise naturally, for which reason people who too often inhibit their natural reflexes are likely to become severely constipated.

In the newborn baby and in some persons with transected spinal cords, the defecation reflexes cause automatic emptying of the lower bowel without the normal control exercised through contraction of the external anal sphincter

OTHER AUTONOMIC REFLEXES AFFECTING BOWEL ACTIVITY

Aside from the duodenocolic, gastrocolic, gastroileal, enterogastric, and defecation reflexes which have been discussed in this chapter, several other important nervous reflexes can affect the overall degree of bowel activity. These are the peritoneointestinal reflex, renointestinal reflex, vesicointestinal reflex, and somatointestinal reflex. All these reflexes are initiated by sensory signals that pass to the prevertebral sympathetic ganglia or to the spinal cord and then are transmitted through the sympathetic nervous system back to the gut. And they all *inhibit* gastrointestinal activity

The *peritoneointestinal reflex* results from irritation of the peritoneum; it strongly inhibits the excitatory enteric nerves and thereby causes intestinal paralysis The *renointestinal* and *vesicointestinal reflexes* inhibit intestinal activity as a result of kidney or bladder irritation. Finally, the *somatointestinal reflex* causes intestinal inhibition when the skin over the abdomen is irritatingly stimulated.

REFERENCES

Atanassova, E , and Papasova. M. Gastrointestinal motility. *Int Rev Physiol*, *12* 35, 1977

Barrett, R. H , and Hanson, M L. Oral Myofunctional Disorders St Louis, C. V. Mosby, 1978

Bofak Gyovai, L A, and Manzione, J V Oral Medicine Patient Evaluation and Management Baltimore, Williams & Wilkins, 1980

Brooks, F P (ed) Gastrointesinal Pathophysiology New York, Oxford University Press, 1978

Bulbring, E (ed) Smooth Muscle An Assessment of Current Knowledge Austin, University of Texas Press, 1981

Chou, C C Relationship between intestinal blood flow and motility *Annu Rev Physiol*, *44* 29 1982

Cohen, S , *et al* Gastrointestinal motility *In* Crane, R K (ed) *International Review of Physiology Gastrointestinal Physiology III* Vol 19 Baltimore, University Park Press, 1979, p. 107

Davenport, H W A Digest of Digestion, 2nd Ed Chicago, Year Book Medical Publishers, 1978

Dickray, G Comparative biochemistry and physiology of gut hormones *Annu Rev Physiol*, *41* 83 1979

Donald, D E Splanchnic circulation *In* Shepherd, J T, and Abboud, F M (eds) *Handbook of Physiology Sec 2 Vol III Bethesda*, American Physiological Society, 1983, p 219

Dubner A , *et al* The Neural Basis of Oral and Facial Function New York, Plenum Press, 1978

Duthie, H S (ed) Gastrointestinal Motility in Health and Disease Baltimore University Press 1978

Furness, J B and Costa, M The adrenergic innervation of the gastrointestinal tract *Erzeb Physiol*, *69* 2, 1974

Gabella G Structural apparatus for force transmission in smooth muscle *Physiol Rev*, *64* 455, 1984

Grossman, M I Neural and hormonal regulation of gastrointestinal function. An overview *Annu Rev Physiol* 41 27 1979

Hunt, J N Mechanisms and disorders of gastric emptying *Annu Rev Med*, *34* 219, 1983

Hurwitz A L, *et al* Disorders of Esophageal Motility Philadelphia, W B Saunders Co, 1979

Jenkins, G N The Physiology and Biochemistry of the Mouth Philadelphia, J B Lippincott Co, 1978

Johnson, L R. Gastrointestinal hormones and their functions. *Annu Rev Physiol*, *39* 135, 1977

Kirsner J B , and Winans, C S The stomach *In* Sodeman, W A , Jr , and Sodeman, T M (eds) Sodeman's Pathologe Physiology, Mechanisms of Disease, 6th Ed Philadelphia, W B Saunders Co , 1979, p 798

Klimov P K Behavior of the organs of the digestive system *Neurosci Behav Physiol*, *14* 333, 1984

Levitt, M D and Bond, J H Flatulence *Annu Rev Med*, *31* 127, 1980

Loewenstein, W R. Junctional intercellular communication The cell-to-cell membrane channel *Physiol Rev*, *61* 829, 1981

Luschei, E S, and Goldberg, L J Neural mechanisms of mandibular control Mastication and voluntary biting *In* Brooks, V B (ed.) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1983, p 1237

Miller A J Deglutition *Physiol Rev*, *62* 129, 1982

Phillips S F, and Devroede G J Functions of the large intestine *In* Crane, R K (ed) International Review of Physiology Gastrointestinal Physiology III Vol 19 Baltimore, University Park Press, 1979, p 263

Rebfeld, J F Gastrointestinal hormones *In* Crane, R K (ed) International Review of Physiology Gastrointestinal Physiology III Vol 19 Baltimore University Park Press, 1979, p 291

Sernka T J, and Jacobson, E D Gastrointestinal Physiology; The Essentials Baltimore, Williams & Wilkins, 1979

Sodeman, W A., Jr, and Watson, D W., The large intestine *In* Sodeman, W A , Jr , and Sodeman, T M (eds.) Sodeman's Pathologic Physiology, Mechanisms of Disease, 6th Ed. Philadelphia, W B Saunders Co , 1979, p 860

Szurszewski, J H Physiology of mammalian prevertebral ganglia *Annu Rev Physiol*, *43* 53, 1981

Van Der Reis, L. (ed.). The Esophagus New York, S Karger, 1978

Weems, W A The intestine as a fluid propelling system *Annu. Rev Physiol*, *43* 9, 1981

Weisbrodt, N W Patterns of intestinal motility *Annu Rev. Physiol*, *43* 21, 1981

Wood, J D. Intrinsic neural control of intestinal motility *Annu Rev Physiol*, *43* 33, 1981.

Young, J A, and vanLennep, E W , The Morphology of Salivary Glands. New York, Academic Press, 1978.

64

Secretory Functions of the Alimentary Tract

Throughout the gastrointestinal tract secretory glands subserve two primary functions. First, digestive enzymes are secreted in most areas from the mouth to the distal end of the ileum. Second, mucous glands, present from the mouth to the anus, provide mucus for lubrication and protection of all parts of the alimentary tract.

Most digestive secretions are formed only in response to the presence of food in the alimentary tract, and the quantity secreted in each segment of the tract is almost exactly the amount needed for proper digestion. Furthermore, in some portions of the gastrointestinal tract even the types of enzymes and other constituents of the secretions are varied in accordance with the types of food present The purpose of the present chapter, therefore, is to describe the different alimentary secretions, their functions, and regulation of their production.

GENERAL PRINCIPLES OF GASTROINTESTINAL SECRETION

ANATOMICAL TYPES OF GLANDS

Several types of glands provide the different types of secretions in the gastrointestinal tract First, on the surface of the epithelium in most parts of the gastrointestinal tract are literally billions of *single cell mucous glands* called simply *mucous cells* or sometimes *goblet cells* These function entirely by themselves without the necessity of coordination with other mucous cells, and they simply extrude their mucus directly into the lumen of the gastrointestinal tract.

Second, many surface areas of the gastrointestinal tract are lined by pits which represent invaginations of the epithelium into the submucosa. In the small intestine these pits, called *crypts of Lieberkühn*, are deep and contain specialized secretory cells One of these will be illustrated later in Figure 64–13. They are lined with a mixture of mucous cells that produce mucus and other epithelial cells that produce mainly serous fluid

[illegible], in the stomach and upper duodenum are found [illegible] numbers of deep *tubular glands*. A typical tubular gland will be illustrated in Figure 64–4, which shows an acid- and pepsinogen-secreting gland of the stomach.

Fourth, also associated with the gastrointestinal tract are several complex glands—the *salivary glands*, the *pancreas*, and the *liver*—which provide secretions for digestion or emulsification of food. The liver has a highly specialized structure that will be discussed in Chapter 70 The salivary glands and the pancreas are compound acinous glands of the type illustrated in Figure 64–2. These glands lie completely outside the walls of the gastrointestinal tract and, in this, differ from all other gastrointestinal glands The major secreting structures, called *acini*, are lined with secreting glandular cells; these feed into a system of ducts that finally empty through one or more portals into the intestinal tract itself

BASIC MECHANISMS OF STIMULATION OF THE GASTROINTESTINAL GLANDS

Effect of Local Stimuli. The mechanical presence of food in a particular segment of the gastrointestinal tract usually causes the glands of that region, and often of adjacent regions, to secrete moderate to large quantities of digestive juices. Part of this local effect results from direct stimulation of the surface glandular cells by contact with the food. For instance, the mucous cells on the surface of the epithelium are stimulated mainly in this way. However, most local stimulation of the intestinal glands results from one of the following three methods of stimulation:

(1) Tactile stimulation or chemical irritation of the mucosa can elicit reflexes that activate the *enteric nervous system* of the intestinal wall to stimulate either the mucous cells on the surface or the deeper glands of the mucosa.

(2) Distension of the gut can also elicit nervous reflexes that stimulate secretion.

(3) Tactile or chemical stimuli or distension can result in increased motility of the gut, as described in the preceding chapter, and the motility in turn can then increase the rate of secretion.

Autonomic Stimulation of Secretion. *Parasympathetic Stimulation.* Stimulation of the par-

asympathetic nerves to the alimentary tract almost invariably increases the rates of glandular secretion. This is especially true of the glands in the upper portion of the tract, innervated by the vagus and other cranial parasympathetic nerves—including the salivary glands, esophageal glands, gastric glands, pancreas, and Brunner's glands in the duodenum—and also of the glands of the distal portion of the large intestine, innervated by the pelvic parasympathetic nerves. Secretion in the remainder of the small intestine and in the first two thirds of the large intestine occurs mainly in response to local stimuli.

Sympathetic Stimulation. Stimulation of sympathetic nerves in some parts of the gastrointestinal tract causes a slight increase in secretion by the respective glands. On the other hand, sympathetic stimulation also results in constriction of the blood vessels supplying the glands. Therefore, sympathetic stimulation can have a dual effect: First, sympathetic stimulation alone can slightly increase secretion. But, second, if parasympathetic or hormonal stimulation is causing copious secretion by the glands, superimposed sympathetic stimulation usually reduces the secretion mainly because of reduced blood supply.

Regulation of Glandular Secretion by Hormones. In the stomach and intestina several different *gastrointestinal hormones* help regulate the volume and character of the secretions. These hormones are liberated from the gastrointestinal mucosa in response to the presence of foods in the lumen of the gut. They then are absorbed into the blood and are carried to glands where they stimulate secretion. This type of stimulation is particularly valuable in increasing the output of gastric juice and pancreatic juice when food enters the stomach or duodenum. Also, hormonal stimulation of the gallbladder wall causes it to empty its stored bile into the duodenum. Other hormones that are still of doubtful value have also been postulated to stimulate secretion by the glands of the small intestine.

Chemically, the gastrointestinal hormones are polypeptides or polypeptide derivatives.

BASIC MECHANISM OF SECRETION BY GLANDULAR CELLS

Secretion of Organic Substances. Though all the basic mechanisms by which glandular cells form different secretions and then extrude these to the exterior are not known, experimental evidence points to the following basic principles of secretion by glandular cells, as illustrated in Figure 64–1.

(1) The nutrient material needed for formation of the secretion must diffuse or be actively transported from the capillary into the base of the glandular cell.

(2) Many *mitochondria* located inside the cell near its base provide oxidative energy for formation of adenosine triphosphate.

(3) Energy from the adenosine triphosphate (see Chapter 67), along with appropriate nutrients, is then used for synthesis of the organic substances; this synthesis occurs almost entirely on the *endoplasmic reticulum*. The *ribosomes* adherent to this reticulum are specifically responsible for formation of proteins that are to be secreted.

(4) The secretory materials are transported into and through the tubules of the endoplasmic reticulum, passing in about 20 minutes all the way to the vesicles of the Golgi complex, which lies near the secretory ends of the cells.

(5) The materials then are modified, added to, concentrated, and discharged into the cytoplasm in the form of *secretory vesicles* that are stored in the apical ends of the secretory cells.

(6) These vesicles remain stored until nervous or hormonal control signals cause them to extrude their contents through the cell's surface. This probably occurs in the following way: The control signal first increases the cell membrane permeability to calcium, and calcium enters the cell. The calcium in turn causes many of the vesicles to fuse with the cell membrane and then to break open on their outer surfaces, thus emptying their contents to the exterior; this process is called *exocytosis*.

Water and Electrolyte Secretion. A second necessity for glandular secretion is sufficient water and electrolytes to be secreted along with the organic substances. The following is a postulated method by which nervous stimulation causes water and salts to pass through the glandular cells in great profusion, which washes the organic substances through the secretory border of the cells at the same time:

(1) Nerve stimulation has a specific effect on the *basal* portion of the cell membrane to cause activa transport of chloride ions to the interior. (2) The resulting increase in electronegativity inside the cell then causes positive ions also to move to the interior of the cell. (3) The excess of both of these ions inside the cell creates an osmotic force which

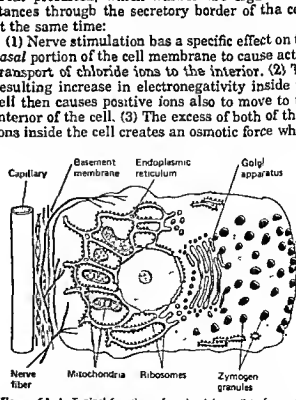


Figure 64–1. Typical function of a glandular cell in formation and secretion of enzymes or other secretory substances.

pulls water to the interior, thereby increasing the hydrostatic pressure inside the cell and causing the cell itself to swell. (4) The pressure in the cell then results in minute ruptures of the secretory border of the cell and causes flushing of water, electrolytes, and organic materials out of the secretory end of the glandular cell and into the lumen of the gland.

In support of this theory have been the following findings. First, the nerve endings on glandular cells are principally on the bases of the cells. Second, microelectrode studies show that the normal electrical potential across the membrane at the base of the cell is between 30 and 40 mv, with negativity on the interior and positivity on the exterior. Parasympathetic stimulation increases this polarization voltage to values some 10 and 20 mv more negative than normal. This increase in polarization occurs a second or more after the nerve signal has arrived, indicating that it is caused by movement of negative ions (presumably chloride ions) through the membrane to the interior of the cell.

Though this mechanism for secretion is still partly theoretical, it does explain how it would be possible for nerve impulses to regulate secretion. Obviously, hormonal effects on the cell membrane could cause similar results.

LUBRICATING AND PROTECTIVE PROPERTIES OF MUCUS AND ITS IMPORTANCE IN THE GASTROINTESTINAL TRACT

Mucus is a thick secretion composed mainly of water, electrolytes, and a mixture of several glycoproteins which themselves are composed of large polysaccharides bound with much smaller quantities of protein. Mucus is slightly different in different parts of the gastrointestinal tract, but everywhere it has several important characteristics that make it both an excellent lubricant and a protectant for the wall of the gut. *First*, mucus has adherent qualities that make it adhere tightly to the food or other particles and also to spread as a thin film over the surfaces. *Second*, it has sufficient *body* that it coats the wall of the gut and prevents actual contact of food particles with the mucosa. *Third*, mucus has a low resistance for slippage so that the particles can slide along the epithelium with great ease. *Fourth*, mucus causes fecal particles to adhere to each other to form the fecal masses that are expelled during a bowel movement. *Fifth*, mucus is strongly resistant to digestion by the gastrointestinal enzymes. And, *sixth*, the glycoproteins of mucus have amphoteric properties and are therefore capable of buffering small amounts of either acids or alkalies; also, mucus often contains moderate quantities of bicarbonate ions, which specifically neutralize acids.

In summary, mucus has the ability to allow easy slippage of food along the gastrointestinal tract and also to prevent excoriative or chemical damage to the epithelium. A person becomes acutely aware of the lubri- qualities of mucus when the salivary glands fail secrete saliva, for under these circumstances it is extremely difficult to swallow solid food even when it is taken with large amounts of water.

SECRETION OF SALIVA

The Salivary Glands; Characteristics of Saliva. The principal glands of salivation are the *parotid*, *submandibular*, and *sublingual* glands; in addition, there are many small *buccal* glands. The daily secretion of saliva normally ranges between about 800 and 1500 milliliters, as shown in Table 64–1.

Saliva contains two major types of protein secretion: (1) a *serous secretion* containing *ptyalin* (an α-amylase), which is an enzyme for digesting starches, and (2) mucous secretion containing *mucin* for lubricating purposes. The parotid glands secrete entirely the serous type, and the submandibular and sublingual glands secrete both the serous type and mucus. The buccal glands secrete only mucus. Saliva has a pH between 6.0 and 7.4, a favorable range for the digestive action of ptyalin.

Secretion of Ions in the Saliva. Saliva contains especially large quantities of potassium and bicarbonate ions. On the other hand, the concentrations of both sodium and chloride ions are considerably less in saliva than in plasma. One can understand these special concentrations of ions in the saliva from the following description of the mechanism for secretion of saliva.

Figure 64–2 illustrates secretion by the submaxillary gland, a typical compound gland containing both *acini* and *salivary ducts*. Salivary secretion is a two-stage operation, the first stage involves the acini and the second, the salivary ducts. The acini secrete a *primary secretion* that contains ptyalin and/or mucin in a solution of ions in concentrations not greatly different from those of typical extracellular fluid. However, as the primary secretion flows through the ducts, two major active transport processes take place that markedly modify the ionic composition of the saliva.

First, *sodium ions* are actively reabsorbed from all the salivary ducts, and *potassium ions* are actively secreted, but at a slower rate, in exchange for the sodium. Therefore, the sodium concentration of the saliva becomes greatly reduced whereas

Table 64–1 DAILY SECRETION OF INTESTINAL JUICES

	Daily Volume (ml)	pH
Saliva	1000	6.0–7.0
Gastric secretion	1500	1.0–3.5
Pancreatic secretion	1000	8.0–8.3
Bile	1000	7.8
Small intestinal secretion	1800	7.5–8.0
Brunner's gland secretion	200	8.0–8.9
Large intestinal secretion	200	7.5–8.0
Total	6700	

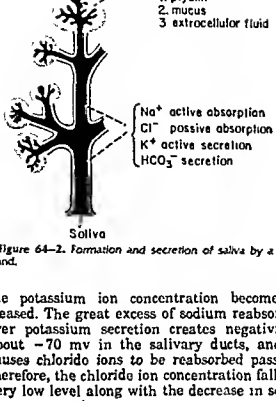


Figure 64–2. Formation and secretion of saliva by a salivary gland.

the potassium ion concentration becomes increased. The great excess of sodium reabsorption over potassium secretion creates negativity of about −70 mv in the salivary ducts, and this causes chloride ions to be reabsorbed passively; therefore, the chloride ion concentration falls to a very low level along with the decrease in sodium ion concentration.

Second, *bicarbonate ions* are secreted by the ductal epithelium into the lumen of the duct. This is probably an active secretory process, though little is known about its details.

The net result of these active transport processes is that, under resting conditions, the concentrations of sodium and chloride ions in the saliva are only about 15 mEq/liter each, approximately 1/7 to 1/10 their concentrations in plasma. On the other hand, the concentration of potassium ions is about 30 mEq/liter, seven times as great as its concentration in plasma, and the concentration of bicarbonate ions is 50 to 70 mEq/liter, about two to three times that of plasma.

During maximal salivation, the salivary ionic concentrations change considerably because the rate of formation of primary secretion by the acini can increase as much as 20-fold. As a result this secretion then flows through the ducts so rapidly that the ductal reconditioning of the secretion is considerably reduced. Therefore, when copious quantities of saliva are secreted, the sodium chloride concentration rises to about one half to two thirds that of plasma, whereas the potassium concentration falls to only four times that of plasma.

In the presence of excess aldosterone secretion, the sodium and chloride reabsorption and the potassium secretion become greatly increased so that the sodium chloride concentration in the saliva is sometimes reduced almost to zero, while the potassium concentration increases still more.

Because of the high potassium ion concentration of saliva, in any abnormal state in which the saliva is lost to the exterior of the body for long periods of time a person can develop serious depletion of potassium ions in the body, leading eventually to serious hypokalemia and paralysis, as was discussed in Chapter 10.

Function of Saliva for Oral Hygiene. Under basal conditions, about 0.5 ml/min of saliva, almost entirely of the mucous type, is secreted all the time except during sleep, when the secretion becomes very little. This secretion plays an exceedingly important role in maintaining healthy oral tissues. The mouth is loaded with pathogenic bacteria that can easily destroy tissues and can also cause dental caries. However, saliva helps prevent the deteriorative processes in several ways. First, the flow of saliva itself helps wash away the pathogenic bacteria, as well as the food particles that provide their metabolic support. Second, the saliva also contains several factors that actually destroy bacteria. One of these is *thiocyanate ions* and another is several *proteolytic enzymes*—most importantly, *lysozyme*—that (a) attack the bacteria, (b) aid the thiocyanate ions in entering the bacteria where they in turn become bactericidal, and (c) digest food particles, thus helping further to remove the bacterial metabolic support. And, third, saliva often contains significant amounts of protein antibodies that can destroy the oral bacteria, including those that cause dental caries.

Therefore, in the absence of salivation, the oral tissues become ulcerated and otherwise infected, and caries of the teeth becomes rampant.

Nervous Regulation of Salivary Secretion. Figure 64–3 illustrates the nervous pathways for regulation of salivation, showing that the salivary glands are controlled mainly by *parasympathetic nervous signals from the salivatory nuclei.* The salivatory nuclei are located approximately at the juncture of the medulla and pons and are excited by both taste and tactile stimuli from the tongue and other areas of the mouth. Many taste stimuli, especially the sour taste, elicit copious secretion of saliva—often as much as 5 to 8 ml per minute or

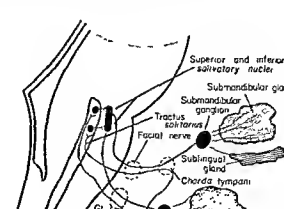


Figure 64–3. Nervous regulation of

8 to 20 times the basal rate of secretion. Also, certain tactile stimuli, such as the presence of smooth objects in the mouth (a pebble, for instance), cause marked salivation, while rough objects cause less salivation and occasionally even inhibit salivation.

Salivation can also be stimulated or inhibited by impulses arriving in the salivatory nuclei from higher centers of the central nervous system. For instance, when a person smells or eats favorite foods, salivation is greater than when disliked food is smelled or eaten. The *appetite area* of the brain that partially regulates these effects is located in close proximity to the parasympathetic centers of the anterior hypothalamus, and it functions to a great extent in response to signals from the taste and smell areas of the cerebral cortex or amygdala

Finally, salivation also occurs in response to reflexes originating in the stomach and upper intestines—particularly when very irritating foods are swallowed or when a person is nauseated because of some gastrointestinal abnormality The swallowed saliva presumably helps to remove the irritating factor in the gastrointestinal tract by diluting or neutralizing the irritant substances.

ESOPHAGEAL SECRETION

The esophageal secretions are entirely mucoid in character and principally provide lubrication for swallowing The main body of the esophagus is lined with many *simple mucous glands*, but at the gastric end, and to a lesser extent in the initial portion of the esophagus, there are many *compound mucous glands*. The mucus secreted by the compound glands in the upper esophagus prevents mucosal excoriation by the newly entering food, whereas the compound glands near the esophagogastric junction protect the esophageal wall from digestion by gastric juices that reflux into the lower esophagus Despite this protection, a peptic ulcer at times may occur at the gastric end of the esophagus

GASTRIC SECRETION

CHARACTERISTICS OF THE GASTRIC SECRETIONS

In addition to the mucus-secreting cells that line the surface of the stomach, the stomach mucosa has two different types of tubular glands: the *oxyntic* (or *gastric*) *glands* and the *pyloric glands* The oxyntic glands, which mean acid-forming glands, secrete *hydrochloric acid, pepsinogen, intrinsic factor*, and mucus, and the pyloric glands secrete mainly *mucus* for protection of the pyloric mucosa but also some *pepsinogen* as well as the hormone *gastrin*. The oxyntic glands are located over the entire wall of the body and fundus of the stomach except along the lesser curvature, and the glands are located in the antral portion of stomach. In addition, a few *cardiac glands*, which are almost identical to the pyloric glands and almost entirely secrete mucus, are located about 1 cm immediately surrounding the entry point of the esophagus.

The Secretions from the Oxyntic Glands. A typical oxyntic gland is shown in Figure 64–4. It is composed of three different types of cells: the *mucous neck cells*, which secrete mainly mucus but also some pepsinogen; the *peptic* (or *chief*) *cells*, which secrete large quantities of pepsinogen; and the *oxyntic* (or *parietal*) *cells*, which secrete hydrochloric acid and intrinsic factor and which lie mainly behind the mucous neck cells or, less often, behind the chief cells A postulated mechanism for secretion of mucus and pepsinogen by the mucous neck cells and the peptic cells was given earlier in the chapter and was depicted in Figure 64–1. However, secretion of hydrochloric acid by the oxyntic (parietal) cells involves particular problems which require further consideration as follows:

Basic Mechanism of Hydrochloric Acid Secretion. The oxyntic (parietal) cells secrete an electrolytic solution containing a maximum of about 160 millimoles of hydrochloric acid per liter, which is almost exactly isotonic with the body fluids. The pH of this acid solution is approximately 0 8, thus illustrating its extreme acidity. At this pH the hydrogen ion concentration is about 3 million times that of the arterial blood. And to concentrate the hydrogen ions this tremendous amount requires over 1500 calories of energy per liter of gastric juice, as discussed in Chapter 9 in relation to membrane transport mechanisms.

Figure 64–5 illustrates the basic structure of an oxyntic cell, showing that it contains a system of *intracellular canaliculi*. The hydrochloric acid is formed at the membranes of these canaliculi and then conducted through openings to the exterior.

Different suggestions for the precise mechanism of hydrochloric acid formation have been offered. One of these is illustrated in Figure 64–6 and consists of the following steps:

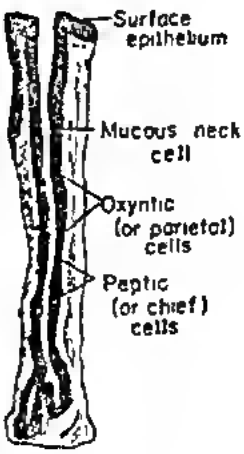


Figure 64–4. An oxyntic gland from the body or fundus of the stomach.

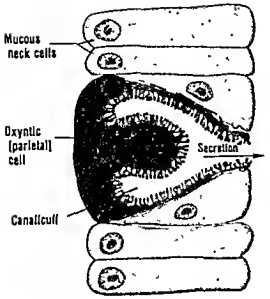


Figure 64–5. Anatomy of the canaliculi in an oxyntic (parietal) cell.

1. Chloride ion is actively transported from the cytoplasm of the oxyntic cell into the lumen of the canaliculus. This creates a negative potential of −40 to −70 millivolts in the conaliculus, which in turn causes passive diffusion of positively charged potassium ions from the cell cytoplasm also into the canaliculus. Thus, in effect, potassium chloride enters the canaliculus.

2. Water is dissociated into hydrogen ions and hydroxyl ions in the cell cytoplasm. The hydrogen ion is then actively secreted into the canaliculus in exchange for potassium ions, this active exchange process being catalyzed by H^+-K^+ ATPase. Thus, most of the potassium ions that had been secreted are reabsorbed, and hydrogen ions take their place in the canaliculus.

3. Water passes through the cell and into the canaliculus by osmosis. Thua the final secretion entering the canaliculus is a solution containing hydrochloric acid in a concentration of 160 millimoles per liter and potassium chloride in a concentration of 17 millimoles per liter.

4. Finally, carbon dioxide, either formed during metabolism in the cell or entering the cell from the blood, combines with water under the influence of carbonic anhydrase to form carbonic acid. This, in turn, dissociates into bicarbonate ion and hydrogen ion. The hydrogen ion combines with the hydroxyl ion released in Step 1 to form water. The bicarbonate ion, in turn, diffuses out of the cell into the blood in exchange for chloride ions that enter the cell and later will be secreted into the canaliculus. The importance of carbon dioxide in the chemical reactions for formation of hydrochloric acid is illustrated by the fact that carbonic anhydrase inhibition by the drug *acetazolamide* almost completely blocks the formation of hydrochloric acid.

Activation of Pepsinogen. Several different types of pepsinogen are secreted by the peptic and mucous cells of the gastric glands. Even so, all the pepsinogens perform essentially the same functions. When the pepsinogens are first secreted, they have no digestive activity. However, as soon as they come in contact with hydrochloric acid, and especially when they come in contact with previously formed pepsin plus the hydrochloric acid, they are immediately activated to form active *pepsin.* In this process, the pepsinogen molecule, having n molecular weight of about 42,500, is split to the pepsin molecule, having a molecular weight of about 35,000.

Pepsin is an active proteolytic enzyme in a highly acid medium (optimum pH = 1.8 to 3.5), but above a pH of about 5 it has little proteolytic activity and soon becomes completely inactivated. Therefore, hydrochloric acid secretion is equally as necessary as pepsin secretion for protein digestion in the stomach; this will be discussed in the following chapter.

Secretion of Other Enzymes. Small quantities of other enzymes are also secreted in the stomach juices, including *gastric lipase, gastric amylase,* and a *gelatinase.* Gastric lipase is of little quantitative importance and is actually a *tributyrase,* for its principal activity is on tributyrin, which is butterfat; it has almost no lipolytic activity on the other fats. Gastric amylase plays a very minor role in digestion of starches, and the gelatinase helps to liquefy some of the proteoglycans in meats.

Secretion of Intrinsic Factor. A substance called *intrinsic factor,* essential for absorption of vitamin B_{12} in the ileum, is secreted by the *oxyntic cells* along with the secretion of hydrochloric acid. Therefore, when the acid-producing cells of the stomach are destroyed, which frequently occurs in chronic gastritis, the person not only develops *achlorhydria* but also develops *pernicious anemia* because of failure of maturation of the red blood cells in the absence of vitamin B_{12} stimulation of the bone marrow. This was discussed in detail in Chapter 4.

Secretion of Mucus in the Stomach. The pyloric glands are structurally similar to the oxyntic glands, but contain few peptic cells and almost no

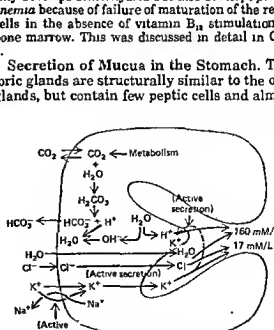


Figure 64–6. Postulated mechanism for the secretion of hydrochloric acid

oxyntic cells. Instead, they contain mostly mucous cells that are identical with the mucous neck cells of the gastric glands These cells secrete a small amount of pepsinogen, as discussed earlier, and an especially large amount of thin mucus that protects the stomach wall from digestion by the gastric enzymes.

In addition, the surface of the stomach mucosa between glands has a continuous layer of mucous cells that secrete large quantities of a far more *viscid and alkaline mucus* that coats the mucosa with a mucous gel layer often more than 1 mm thick, thus providing a major shell of protection for the stomach wall as well as contributing to lubrication of food transport. Even the slightest irritotion of the mucosa directly stimulates the mucous cells to secrete copious quantities of this thick, viscid mucus. This in turn forms a gastric barrier that prevents digestion of the gastric wall and also greatly reduces the absorption of most substances by the gastric mucosa.

REGULATION OF GASTRIC SECRETION BY NERVOUS AND HORMONAL MECHANISMS

Gastric secretion is regulated by both nervous and hormonal mechanisms; nervous regulation is effected through the parasympathetic fibers of the vagus nerves as well as through local enteric nervous system reflexes, and hormonal regulation takes place mainly in response to the hormone *gastrin*. Thus, regulation of gastric secretion is different from the regulation of salivary secretion, which is effected entirely by nervous mechanisms

Vagal Stimulation of Gastric Secretion

Nervous signals to cause gastric secretion originate in the dorsal motor nuclei of the vagi and pass via the vagus nerves to the enteric nervous system of the stomach wall and thence to the oxyntic glands In response, these glands secrete vast quantities of both pepsin and acid, but with a higher proportion of pepsin than in gastric juice elicited in other ways. Also, vagal signals to all the mucus-secreting glands and epithelial lining cells cause increased mucus secretion as well.

Still another effect of vagal stimulation is to cause the antral part of the stomach mucosa to secrete the hormone *gastrin* As will be explained in the following paragraphs, this hormone then acts on the gastric glands to cause additional flow of highly acidic gastric juice. Thus, vagal stimulation excites stomach secretion both directly by stimulation of the gastric glands and indirectly through the gastrin mechanism.

Stimulation of Gastric Secretion by Gastrin

When food enters the stomach, it causes the portion of the stomach mucosa to secrete hormone gastrin. This hormone is secreted by cells, also called *G cells*, in the pyloric glands and to a lesser extent in the proximal glands (Brunner's glands) of the duodenum. Gastrin is a large peptide secreted in two forms, a large form called *G-34*, containing 34 amino acids, and a smaller form, *G-17*, containing 17 amino acids Though both of these are important, the smaller form is more abundant.

The food causes release of this hormone in two ways: (1) The actual bulk of the food distends the stomach, and this causes the hormone gastrin to be released from the antral mucosa. (2) Certain substances called secretagogues—such as food extractives, partially digested proteins, alcohol (in low concentration), caffeine, and so forth—also cause gastrin to be liberated from the antral mucosa.

Both of these stimuli—the distension and the chemical action of the secretagogues—elicit gastrin release by means of a local nerve reflex. That is, they stimulate sensory nerve fibers in the stomach epithelium which in turn synapse with the enteric nervous system. This then transmits efferent signals to the gastrin cells, causing them to secrete the gastrin Therefore, any factor that blocks this reflex will also block the formation of gastrin. For instance, anesthetization of the gastric mucosa to block the sensory stimuli will prevent gastrin release; administration of atropine, which blocks the action on the gastrin cells of the acetylcholine released by the enteric nerves, will also prevent gastrin release

Gastrin is absorbed into the blood and carried to the oxyntic glands in the body of the stomach where it stimulates mainly the oxyntic cells but to a lesser extent the peptic cells, also. The oxyntic cells increase their rate of hydrochloric acid secretion as much as eightfold, and the peptic cells increase their rate of enzyme secretion two- to fourfold

The rate of secretion in response to gastrin is somewhat less than to vagal stimulation, 200 ml per hour in contrast to about 500 ml per hour, indicating that the gastrin mechanism is a less potent acute mechanism for stimulation of stomach secretion than is vagal stimulation. However, the gastrin mechanism usually continues for several hours in contrast to a much shorter period of time for vagal stimulation. Therefore, as a whole, it is likely that the gastrin mechanism is equally as important as, if not more important than, the vagal mechanism for control of gastric secretion. Yet, when both of these work together, the total secretion is much greater than the sum of the individual secretions caused by each of the two mechanisms In other words, *the two mechanisms multiply each other rather than simply add to each other.*

Role of Histamine in Controlling Gastric Secretion

Histamine, an amino acid derivative, also stimulates gastric secretion. When the H-2 type of

histamine receptors (but not the H-1 receptors) of the oxyntic glands are blocked with the histamine antagonist drug *cimetidine*, not only does this block the stimulatory effect of histamine on gastric secretion but the effect of gastrin as well. Therefore, it is believed that local gastric mucosal histamine is a prerequisite for the stimulatory function of gastrin—perhaps gastrin causes the release of histamine and it is the histamine that actually causes the secretion.

Chemical Composition of Gastrin and Other Gastrointestinal Hormones

Below are illustrated the amino acid compositions of *gastrin-17* and also of *cholecystokinin* and *secretin*, which will be discussed later in the chapter. Note that all are polypeptides and that the terminal five amino acids in the gastrin and cholecystokinin molecular chains are exactly the same. The activity of gastrin resides in the terminal four amino acids and in the terminal eight amino acids for cholecystokinin; all the amino acids are essential in secretin. A synthetic gastrin, composed of the terminal four amino acids of natural gastrins plus the amino acid alanine, has all the same physiological properties as the natural gastrins. This synthetic product is called *pentagastrin*.

Feedback Inhibition of Gastric Acid Secretion. When the acidity of the gastric juices increases to a pH of 2.0, the gastrin mechanism for stimulating gastric secretion becomes totally blocked. This effect probably results from two different factors. First, greatly enhanced acidity depresses or blocks the extraction of gastrin itself from the antral mucosa. Second, the acid seems to cause an inhibitory nervous reflex that inhibits gastric acid secretion.

Obviously, this feedback inhibition of the gastric glands plays an important role in protecting the stomach against excessively acid secretions, which would readily cause peptic ulceration. In addition to this protective effect, the feedback mechanism is also important in maintaining optimal pH for function of the peptic enzymes in the digestive process because whenever the pH rises above 2.5 to 3.5 gastrin begins to be secreted again and more acid is secreted.

The Three Phases of Gastric Secretion

Gastric secretion is said to occur in three separate phases (as illustrated in Fig. 64–7): a *cephalic phase*, a *gastric phase*, and an *intestinal phase*. However, as will be apparent in the following discussion, these three phases in reality fuse together.

The Cephalic Phase. The cephalic phase of gastric secretion occurs even before food enters the stomach. It results from the sight, smell, thought, or taste of food; and the greater the appetite, the more intense is the stimulation. Neurogenic signals causing the cephalic phase of secretion can originate in the cerebral cortex or in the appetite centers of the amygdala or hypothalamus. They are transmitted through the dorsal motor nuclei of the vagi to the stomach. This phase of secretion probably accounts for less than one-tenth of the gastric secretion normally associated with eating a meal.

The Gastric Phase. Once the food enters the stomach, it excites the gastrin mechanism, which in turn causes secretion of gastric juice that continues throughout the several hours that the food remains in the stomach

In addition, the presence of food in the stomach also causes (a) local reflexes in the enteric nervous system of the stomach and (b) vagovagal reflexes that pass all the way to the brain stem and back to the stomach; both these reflexes cause acetylcholine stimulation of the oxyntic glands and add to the secretion caused by the gastrin mechanism

The gastric phase of secretion accounts for at least two thirds of the total gastric secretion associated with eating a meal and, therefore, accounts for most of the total daily gastric secretion of about 1500 ml.

The Intestinal Phase. Presence of food in the upper portion of the small intestine, particularly in the duodenum, can cause the stomach to secrete small amounts of gastric juice. This probably results partly from the fact that small amounts of gastrin are also released by the duodenal mucosa in response to distension or chemical stimuli of the same type as those that stimulate the stomach gastrin mechanism It is probable that several other hormones or reflexes also play minor roles in causing secretion of gastric juice.

However, several intestinal factors can also inhibit gastric secretion, and these often completely override the excitatory factors, as we shall discuss in the following paragraph.

Gastrin:

Glu- Gly- Pro- Trp- Leu- Glu- Glu- Glu- Glu- Glu- Ala- Tyr- Gly- Trp- Met- Asp- Phe- NH_2
$\qquad\qquad\qquad

Inhibition of Gastric Secretion by Intestinal Factors

Though chyme stimulates gastric secretion during the intestinal phase of secretion, it paradoxically often inhibits secretion during the gastric phase. This results from at least two different influences:

1. The presence of food in the small intestine initiates an *enterogastric reflex*, transmitted through the enteric nervous system as well as through the extrinsic sympathetic and vagus nerves, that inhibits stomach secretion. This reflex can be initiated by distension of the small bowel, the presence of acid in the upper intestine, the presence of protein breakdown products, or irritation of the mucosa. This is part of the complex mechanism discussed in the preceding chapter for slowing down stomach emptying when the intestines are already filled.

2. The presence of acid, fat, protein breakdown products, hyper- or hypo-osmotic fluids, or any irritating factor in the upper small intestine causes the release of several intestinal hormones. Two of these are *secretin* and *cholecystokinin*. Both of these are especially important for control of pancreatic secretion, and cholecystokinin is also important in promoting emptying of the gallbladder. However, in addition to these effects, they both oppose the stimulatory effects of gastrin. Also, three other hormones, *gastric inhibitory peptide, vasoactive intestinal polypeptide,* and *somatostatin* all have slight to moderate effects on inhibiting gastric secretion.

The functional purpose of the inhibition of gastric secretion by intestinal factors is probably to slow the release of chyme from the stomach when the small intestine is already filled. In fact, the enterogastric reflex and also these inhibitory hormones usually reduce stomach motility at the same time that they reduce gastric secretion, as was discussed in the previous chapter.

Secretion During the Interdigestive Period. The stomach secretes only a few milliliters of gastric juice per hour during the "interdigestive period" when little or no digestion is occurring anywhere in the gut. Furthermore, the secretion that does occur is almost entirely of the so-called nonoxyntic type, meaning that it is composed mainly of mucus containing very little pepsin and almost no acid. In fact, on occasion, the fluid is actually slightly alkaline, containing moderate quantities of sodium bicarbonate. However, strong emotional stimuli frequently increase the interdigestive secretion to 50 ml or more of highly peptic and highly acidic gastric juice per hour, in very much the same manner that the cephalic phase of gastric secretion excites secretion at the onset of a meal. This increase of secretion during the presence of emotional stimuli is believed to be one of the factors in the development of peptic ulcers, as will be discussed in Chapter 66.

PANCREATIC SECRETION

The pancreas, which lies parallel to and beneath the stomach, is a large compound gland with an internal structure almost identical to that of the salivary glands. Digestive enzymes are secreted by the acini, and large volumes of sodium bicarbonate solution are secreted by the small ductules leading from the acini. The combined product then flows through a long pancreatic duct that usually joins the hepatic duct immediately before it empties into the duodenum through the sphincter of Oddi. However, on occasion this duct empties separately into the duodenum. Pancreatic juice is secreted mainly in response to the presence of chyme in the upper portions of the small intestine, and the characteristics of the pancreatic juice are determined to some extent by the types of food in the chyme.

Characteristics of Pancreatic Juice. Pancreatic juice contains enzymes for digesting all three major types of food: proteins, carbohydrates, and fats. It also contains large quantities of bicarbonate ions which play an important role in neutralizing the acid chyme emptied by the stomach into the duodenum.

The proteolytic enzymes are *trypsin, chymotrypsin, carboxypolypeptidase, ribonuclease,* and *deoxyribonuclease*. By far the most abundant of these is trypsin. The trypsin and chymotrypsin split whole and partially digested proteins into peptides of

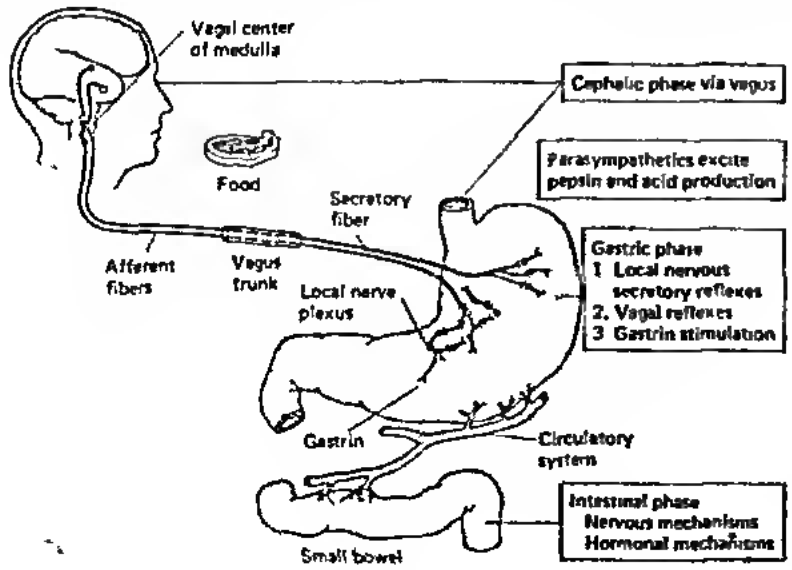


Figure 64–7. The phases of gastric secretion and their regulation.

various sizes, but these do not cause the release of individual amino acids. On the other hand, carboxypolypeptidase splits individual amino acids from the carboxyl ends of the peptides, thus completing the digestion of much of the proteins all the way to the amino acid state. The nucleases split the two types of nucleic acids: ribonucleic and deoxyribonucleic acids.

The digestive enzyme for carbohydrates is *pancreotic amylase*, which hydrolyzes starches, glycogen, and most other carbohydrates except cellulose to form disaccharides and a few trisaccharides.

The main enzymes for fat digestion are *pancreotic lipase*, which is capable of hydrolyzing neutral fat into fatty acids and monoglycerides, *cholesterol esterase*, which causes hydrolysis of cholesterol esters, and *phospholipase*, which splits fatty acids from phospholipids.

When synthesized in the pancreatic cells, the proteolytic enzymes are in the inactive forms *trypsinogen*, *chymotrypsinogen*, and *procarboxypolypeptidase*, which are all enzymatically inactive. These become activated only after they are secreted into the intestinal tract. Trypsinogen is activated by an enzyme called *enterokinase*, which is secreted by the intestinal mucosa when chyme comes in contact with the mucosa. Also, trypsinogen can be autocatalytically activated by trypsin that has already been formed. Chymotrypsinogen is activated by trypsin to form chymotrypsin, and procarboxypolypeptidase is activated in a similar manner.

Secretion of Trypsin Inhibitor. It is important that the proteolytic enzymes of the pancreatic juice not become activated until they have been secreted into the intestine, for the trypsin and the other enzymes would digest the pancreas itself. Fortunately, the same cells that secrete the proteolytic enzymes into the acini of the pancreas secrete simultaneously another substance called *trypsin inhibitor*. This substance is stored in the cytoplasm of the glandular cells surrounding the enzyme granules, and it prevents activation of trypsin both inside the secretory cells and in the acini and ducts of the pancreas. Since it is trypsin that activates the other pancreatic proteolytic enzymes, trypsin inhibitor also prevents the subsequent activation of all these.

However, when the pancreas becomes severely damaged or when a duct becomes blocked, large quantities of pancreatic secretion become pooled in the damaged areas of the pancreas. Under these conditions, the effect of trypsin inhibitor is sometimes overwhelmed, in which case the pancreatic secretions rapidly become activated and literally digest the entire pancreas within a few hours, giving rise to the condition called *acute pancreatitis*. This often is lethal because of accompanying shock, and even if not lethal it leads to a lifetime of pancreatic insufficiency.

Secretion of Bicarbonate Ions. The enzymes of the pancreatic juice are secreted entirely by the acini of the pancreatic glands. On the other hand, two other important components of pancreatic juice, bicarbonate ions and water, are secreted in large amounts mainly by the epithelial cells of the small ductules leading from the acini. We shall see later that the stimulatory mechanisms for (a) enzyme production and (b) production of bicarbonate ions and water are also quite different. When the pancreas is stimulated to secrete copious quantities of pancreatic juice—that is, copious quantities of the bicarbonate ions and water—the bicarbonate ion concentration can rise to as high as 145 mEq/liter, a value approximately five times that of bicarbonate ions in the plasma. Obviously, this provides a large quantity of alkaline ion in the pancreatic juice that serves to neutralize acid emptied into the duodenum from the stomach.

The cellular mechanism for secreting sodium bicarbonate solution into the pancreatic ductules is illustrated in Figure 64–8. The major steps in this process are:

1. Carbon dioxide diffuses to the interior of the cell from the blood and combines with water under the influence of carbonic anhydrase to form carbonic acid. The carbonic acid in turn dissociates into bicarbonate ion and hydrogen ion. Then the bicarbonate ion is actively transported through the *luminal* border of the cell into the lumen of the duct.

2. The hydrogen ion formed by dissociation of carbonic acid inside the cell is exchanged for sodium ion through the *blood* border of the cell, also by an active transport process. The sodium ion in turn either diffuses through or is actively transported through the *luminal* border into the pancreatic duct to provide electrical neutrality for the bicarbonate ion.

3. The movement of sodium and bicarbonate ions from the blood to the lumen creates an osmotic gradient that causes osmosis of water also into the

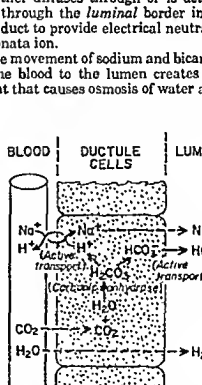


Figure 64–8. Secretion of iso-osmotic sodium bicarbonate solution by the pancreatic ductules.

pancreatic duct, thus forming the bicarbonate solution.

REGULATION OF PANCREATIC SECRETION

Pancreatic secretion, like gastric secretion, is regulated by both nervous and hormonal mechanisms. However, in this case, hormonal regulation is by far the more important.

Nervous Regulation. When the cephalic and gastric phases of stomach secretion occur, parasympathetic impulses are simultaneously transmitted along the vagus nerves to the pancreas, resulting in *acetylcholine* release followed by secretion of moderate to large amounts of enzymes into the pancreatic acini. However, little secretion flows through the pancreatic ducts to the intestine because only small amounts of water and electrolytes are secreted along with the enzymes. Therefore, most of the enzymes are temporarily stored in the acini.

Hormonal Regulation. After chyme enters the small intestine, pancreatic secretion becomes copious, mainly in response to the hormone *secretin*. In addition, a second hormone, *cholecystokinin*, causes still far more increase in the secretion of enzymes.

Stimulation of Secretion of Copious Quantities of Bicarbonate by Secretin—Neutralization of the Acidic Chyme. Secretin is a polypeptide containing 27 amino acids (molecular weight about 3400) that is present in so-called *S cells* in the mucosa of the upper small intestine (duodenum and upper jejunum) in an inactive form *prosecretin*. When acid chyme enters the intestine, it causes the release and activation of secretin, which is subsequently absorbed into the blood. The one constituent of chyme that causes greatest secretin release is hydrochloric acid, though almost any type of food will cause at least some release.

Secretin causes the pancreas to secrete large quantities of fluid containing a high concentration of bicarbonate ion (up to 145 mEq/liter) but a low concentration of chloride ion. However, this fluid contains very little enzymes when the pancreas is stimulated only by secretin because secretin does not stimulate the acinar cells.

The secretin mechanism is especially important for two reasons. *First*, secretin is released from the mucosa of the small intestine when the pH of the duodenal contents falls below 4.5, and its release increases very greatly as the pH falls to 3.0 and as more and more acid reaches greater areas of the duodenum and jejunum. This immediately causes large quantities of pancreatic juice containing abundant amounts of sodium bicarbonate to be secreted, which results in the following reaction in the duodenum:

$$HCl + NaHCO_3 \rightarrow NaCl + H_2CO_3$$

The carbonic acid immediately dissociates into carbon dioxide and water, and the carbon dioxide is absorbed into the blood and expired through the lungs, thus leaving a neutral solution of sodium chloride in the duodenum. In this way, the acid contents emptied into the duodenum from the stomach become neutralized, and the peptic activity of the gastric juices is immediately blocked. Since the mucosa of the small intestine cannot withstand the digestive action of acid gastric juice, this is a highly important and even essential protective mechanism against the development of duodenal ulcers, which will be discussed in further detail in Chapter 66.

A *second* importance of bicarbonate secretion by the pancreas is to provide an appropriate pH for action of the pancreatic enzymes. All these function optimally in a slightly alkaline or neutral medium. The pH of the sodium bicarbonate secretion averages 8.0.

Cholecystokinin—Control of Enzyme Secretion by the Pancreas. The presence of food in the upper small intestine also causes a second hormone, cholecystokinin, a polypeptide containing 33 amino acids, to be released from the mucosa. This results especially from the presence of proteoses and peptones, which are products of partial protein digestion, and of long chain fatty acids; however, acids will also cause its release in smaller quantities. Cholecystokinin, like secretin, passes by way of the blood to the pancreas but, instead of causing sodium bicarbonate secretion, causes secretion of large quantities of digestive enzymes by the acinar cells, which is similar to the effect of vagal stimulation but even more pronounced.

Still another factor with a moderate effect on promoting secretion of enzymes by the pancreatic acini is the hormone *gastrin*. This occurs principally during the gastric phase of gastric secretion when the large amounts of gastrin produced have a moderate stimulating effect similar to that of

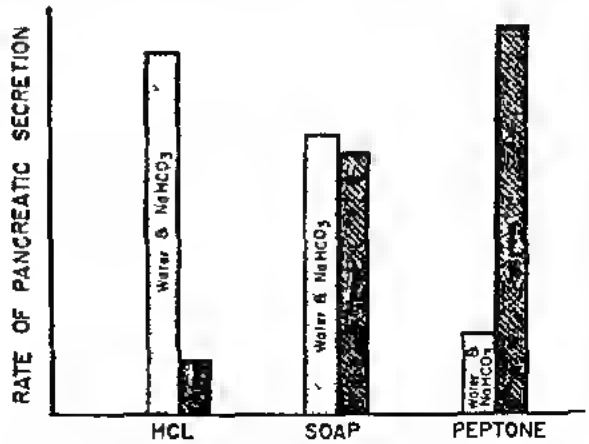


Figure 64–9. Sodium bicarbonate and enzyme secretion by the pancreas, caused respectively by the presence of acid, fat (soap), or peptone solutions in the duodenum.

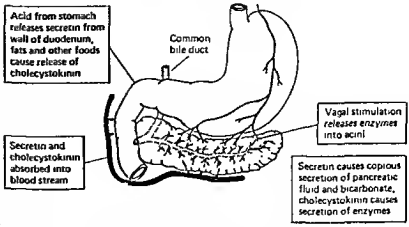


Figure 64–10. Regulation of pancreatic secretion.

cholecystokinin. The cholecystokinin-like effect of gastrin results from the fact that gastrin and cholecystokinin both have the same five terminal amino acids where most of the activity of these hormones resides.

The differences between the stimulatory effects of secretin and cholecystokinin are shown in Figure 64–9, which illustrates (a) intense sodium bicarbonate secretion in response to acid in the duodenum, (b) a dual effect in response to soap (a fat), and (c) intense enzyme secretion in response to peptones.

Figure 64–10 summarizes the overall regulation of pancreatic secretion. The total amount secreted each day is about 1200 ml.

SECRETION OF BILE AND FUNCTIONS OF THE BILIARY TREE

PHYSIOLOGIC ANATOMY OF BILIARY SECRETION

All the hepatic cells continually form a small amount of secretion called *bile*. This is secreted into the minute *bile canaliculi* that lie between the hepatic cells in the hepatic plates, and the bile then flows peripherally toward the interlobular septa where the canaliculi empty into *terminal bile ducts*, then into progressively larger ducts, finally reaching the *hepatic duct* and *common bile duct* from which the bile either empties directly into the duodenum or is diverted through the *cystic duct* into the gallbladder, illustrated in Figure 64–11.

Storage of Bile in the Gallbladder. The bile secreted continually by the liver cells is normally stored in the gallbladder until needed in the duodenum. The total secretion of bile each day is some 700 to 1200 ml, and the maximum volume of the gallbladder is only 30 to 60 ml. Nevertheless, as much as 12 hours' bile secretion can be stored in the gallbladder because water, sodium, chloride, and most other small electrolytes are continually absorbed by the gallbladder mucosa, concentrating the other bile constituents, including the bile salts, cholesterol, lecithin, and bilirubin. Most of this absorption is caused by active transport of sodium through the gallbladder epithelium. Bile is normally concentrated about fivefold, but it can be concentrated up to a maximum of 12- to 18-fold.

Emptying of the Gallbladder. Two basic conditions are necessary for the gallbladder to empty:

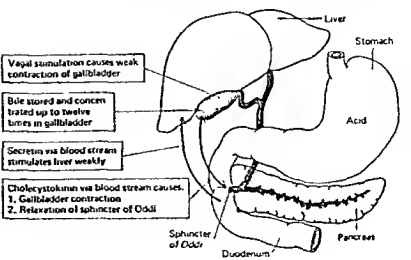


Figure 64–11. Mechanisms of liver secretion and gallbladder emptying.

(1) The sphincter of Oddi must relax to allow bile to flow from the common bile duct into the duodenum, and (2) the gallbladder itself must contract to provide the force required to move the bile along the common duct. After a meal, particularly one that contains a high concentration of fat, both these effects take place in the following manner:

First, the fat (also partially digested protein) in the food entering the small intestine causes release of a hormone called *cholecystokinin* from the intestinal mucosa, especially from the upper regions of the small intestine. The *cholecystokinin* in turn is absorbed into the blood and, on passing to the gallbladder, causes specific contraction of the gallbladder muscle. This provides the pressure that forces bile toward the duodenum.

Second, vagal stimulation associated with the cephalic phase of gastric secretion or with various intestinointestinal reflexes causes an additional weak contraction of the gallbladder.

Third, when the gallbladder contracts, the sphincter of Oddi becomes inhibited, this effect probably resulting from either a neurogenic or a myogenic reflex from the gallbladder to the sphincter of Oddi. This inhibition may also, to some extent, be a direct effect of cholecystokinin on the sphincter, causing relaxation.

Fourth, the presence of food in the duodenum causes the degree of peristalsis in the duodenal wall to increase. Each time a peristaltic wave travels toward the sphincter of Oddi, this sphincter, along with the adjacent intestinal wall, momentarily relaxes because of the phenomenon of "receptive relaxation" that travels ahead of the peristaltic contraction wave, and, if the bile in the common bile duct is under sufficient pressure, a small quantity of the bile squirts into the duodenum.

In summary, the gallbladder empties its store of concentrated bile into the duodenum mainly in response to the cholecystokinin stimulus. When fat is not in the meal, the gallbladder empties poorly, but, when adequate quantities of fat are present, the gallbladder empties completely in about one hour.

Figure 64–11 summarizes the secretion of bile, its storage in the gallbladder, and its release from the bladder to the gut.

Composition of Bile. Table 64–2 gives the composition of bile when it is first secreted by the liver and then after it has been *concentrated* in the gallbladder. This table shows that by far the most abundant substance secreted in the bile is the *bile salts*, accounting for about half of the total solutes of bile, but also secreted or excreted in large concentrations are *bilirubin*, *cholesterol*, *lecithin*, and the usual *electrolytes* of plasma. In the concentrating process in the gallbladder, water and large portions of the electrolytes (except calcium ions) are reabsorbed by the gallbladder mucosa, but essentially all the other constituents, including especially the bile salts and the lipid substances cholesterol and lecithin, are not reabsorbed and therefore become highly concentrated in the gallbladder bile.

Table 64–2 COMPOSITION OF BILE

	Liver Bile	Gallbladder Bile
Water	97.5 gm %	92 gm %
Bile salts	1.1 gm %	6 gm %
Bilirubin	0.04 gm %	0.3 gm %
Cholesterol	0.1 gm %	0.3 to 0.9 gm %
Fatty acids	0.12 gm %	0.3 to 1.2 gm %
Lecithin	0.04 gm %	0.3 gm %
Na^+	145 mEq/l	130 mEq/l
K^+	5 mEq/l	12 mEq/l
Ca^{++}	5 mEq/l	23 mEq/l
Cl^-	100 mEq/l	25 mEq/l
HCO_3^-	28 mEq/l	10 mEq/l

THE BILE SALTS AND THEIR FUNCTION

The liver cells form about 0.5 gram of *bile salts* daily. The precursor of the bile salts is *cholesterol*, which is either supplied in the diet or synthesized in the liver cells during the course of fat metabolism and then converted to *cholic acid* or *chenodeoxycholic acid* in about equal quantities. These acids then combine principally with glycine and to a lesser extent with taurine to form *glyco-* and *tauro-conjugated bile acids*. The salts of these acids are secreted in the bile.

The bile salts have two important actions in the intestinal tract. First, they have a detergent action on the fat particles in the food, which decreases the surface tension of the particles and allows the agitation in the intestinal tract to break the fat globules into minute sizes. This is called the *emulsifying* or *detergent function* of bile salts. Second, and even more important than the emulsifying function, bile salts help in the absorption of fatty acids, monoglycerides, cholesterol, and other lipids from the intestinal tract. They do this by forming minute complexes with these lipids; the complexes are called *micelles*, and they are highly soluble because of the electrical charges of the bile salts. The lipids are "ferried" in this form to the mucosa where they are then absorbed; this mechanism is described in detail in the following chapter. Without the presence of bile salts in the intestinal tract, up to 40 per cent of the lipids are lost into the *stools*, and the person often develops a *metabolic* deficit due to this nutrient loss.

Also, when fats are not absorbed adequately, the fat-soluble vitamins A, D, E, and K are not absorbed satisfactorily. Though large quantities of the first three of these vitamins are usually stored in the body, this is not true of vitamin K. Within only a few days after bile secretion ceases, the person usually develops a deficiency of vitamin K. This in turn results in deficient formation by the liver of several blood coagulation factors—pro-

thrombin, and factors VII, IX, and X—thus resulting in serious impairment of blood coagulation.

Enterohepatic Circulation of Bile Salts. Approximately 94 per cent of the bile salts are reabsorbed by an active transport process through the intestinal mucosa in the distal ileum. They enter the portal blood and pass to the liver. On reaching the liver the bile salts are absorbed almost totally on the first passage through the venous sinusoids into the hepatic cells and then resecreted into the bile. In this way about 94 per cent of all the bile salts are recirculated into the bile, so that on the average these salts make the entire circuit some 18 times before being carried out in the feces. The small quantities of bile salts lost into the feces are replaced by new amounts formed continually by the liver cells This recirculation of the bile salts is called the *enterohepatic circulation*.

The quantity of bile secreted by the liver each day is highly dependent on the availability of bile salts—the greater the quantity of bile salts in the enterohepatic circulation (usually a total of about 2.5 grams), the greater the rate of bile secretion. Indeed, ingestion of an excess of bile salts can increase bile secretion by several hundred milliliters per day When a bile fistula forms so that bile is lost directly from the common bile duct to the exterior, the bile salts cannot be reabsorbed Therefore, the total quantity of bile salts in the enterohepatic circulation becomes greatly depressed, and concurrently the volume of liver secretion is also depressed

However, if a bile fistula continues to empty the bile salts to the exterior for several days to several weeks, the liver increases its production of bile salts as much as tenfold, which increases the rate of bile secretion approximately back to normal. This also demonstrates that the daily rate of bile salt secretion is actively controlled by the availability (or lack of availability) of bile salts in the enterohepatic circulation.

One of the most important functional characteristics of the bile acids is that they normally exist in the form of strongly ionized salts. Because of this ionization, these salts are not absorbed to a significant extent anywhere except in the distal ileum. Here a specific receptor for the bile salts occurs in the mucosal epithelial cells, which promotes specific "pumping" of the bile salts back into the blood.

Role of Secretin in Controlling Bile Secretion. In addition to the strong stimulating effect of bile acids on bile secretion, the hormone *secretin* also increases bile secretion, sometimes more than doubling the secretion rate for several hours after a meal. However, this increase in secretion represents mainly secretion of a bicarbonate-rich watery solution by the epithelial cells of the bile ductules and ducts and not increased secretion of bile acids by the liver parenchymal cells. The bicarbonate in turn passes into the small intestine and joins the bicarbonate from the pancreas in neutralizing the acid from the stomach. Thus, the secretin feedback mechanism for neutralizing duodenal acid operates not only through its effects on pancreatic secretion but through its effect on the secretion by the liver ductules and ducts as well.

SECRETION OF CHOLESTEROL; GALLSTONE FORMATION

Bile salts are formed in the hepatic cells from cholesterol, and in the process of secreting the bile salts about one tenth as much cholesterol is also secreted into the bile. This amounts to a total of 1 to 2 grams a day. No specific function is known for the cholesterol in the bile, and it is presumed that it is simply a byproduct of bile salt formation and secretion.

Cholesterol is almost insoluble in pure water, but the bile salts and lecithin in bile combine physically with the cholesterol to form *ultramicroscopic micelles* that are soluble, as will be explained in more detail in the following chapter. When the bile becomes concentrated in the gallbladder, the bile salts and lecithin become concentrated along with the cholesterol, which keeps the cholesterol in solution. Also, the hormone *secretin*, which is released into the blood from the duodenal mucosa in response to acid in the chyme, causes the bile ducts to secrete about 200 ml of highly alkaline fluid each day. This alkalinity is essential for formation of the micelles, for they will not develop in an acid medium.

Nevertheless, under abnormal conditions the cholesterol may precipitate, resulting in the formation of *gallstones*, as shown in Figure 64–12 The different conditions that can cause cholesterol precipitation are (1) too much absorption of water from the bile, (2) too much absorption of bile salts and lecithin from the bile, (3) too much secretion of cholesterol in the bile, and (4) inflammation of the epithelium of the gallbladder. The latter two of these require the following special explanation.

The amount of cholesterol in the bile is determined partly by the quantity of fat that the person eats, for the hepatic cells synthesize cholesterol as one of the products of fat metabolism in the body. For this reason, persons on a high fat diet over e period of many years are prone to the development of gallstones.

Inflammation of the gallbladder epithelium often results from low grade chronic infection, this changes the absorptive characteristics of the gallbladder mucosa, sometimes allowing excessive absorption of water, bile salts, or other substances that are necessary to keep the cholesterol in solution As a result, cholesterol begins to precipitate, usually forming many small crystals of cholesterol on the surface of the inflamed mucosa. These, in turn, act as nidi for further precipitation of cholesterol, and the crystals grow larger and larger Occasion-

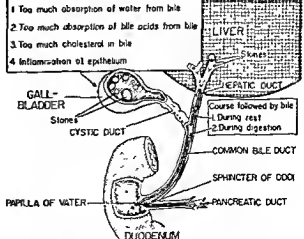


Figure 64–12. Formation of gallstones.

ally tremendous numbers of sandlike stones develop, but much more frequently these coalesce to form a few large gallstones, or even a single stone that fills the entire gallbladder. Also, calcium ions, which are usually concentrated five- or more fold in the gallbladder, often precipitate in the gallstones, making them x-ray–opaque so that they can be seen in radiographs of the abdomen.

Medical Theropy for Dissolving Gallstones. Simple cholesterol gallstones in most patients can be dissolved over a period of one to two years by feeding the patient 1 to 1.5 grams of *chenodeoxycholic acid* daily. This is one of the naturally secreted bile acids, and its exogenous administration adds greatly to the enterohepatic pool of bile acids. This causes dissolution and reabsorption of the gallstones in the following ways: (1) The increased quantity of bile acids increases the volume of bile formed and therefore decreases the concentration of cholesterol in the bile. (2) The increase of bile acids in the bile makes the cholesterol that is present more soluble. (3) The exogenous administration of bile acids decreases the formation of bile acids by the liver, which at the same time reduces the formation of cholesterol.

SECRETIONS OF THE SMALL INTESTINE

SECRETION OF MUCUS BY BRUNNER'S GLANDS AND BY MUCOUS CELLS OF THE INTESTINAL SURFACE

An extensive array of compound mucous glands, called *Brunner's glands,* is located in the first few centimeters of the duodenum, mainly between the pylorus and the papilla of Voter where the pancreatic juices and bile empty into the duodenum. These glands secrete mucus in response to (a) direct tactile stimuli or irritating stimuli of the overlying mucosa, (b) vagal stimulation, which causes secretion concurrently with increase in stomach secretion, and (c) gastrointestinal hormones, especially secretin. The function of the mucus secreted by Brunner's glands is to protect the duodenal wall from digestion by the gastric juice, and their rapid and intense response to irritating stimuli is especially geared to this purpose.

Brunner's glands are inhibited by sympathetic stimulation; therefore, such stimulation is likely to leave the duodenal bulb unprotected and is perhaps one of the factors that cause this area of the gastrointestinal tract to be the site of peptic ulcers in about 50 per cent of the cases.

Mucus is also secreted in large quantities by mucous cells located extensively over the surface of the intestinal mucosa throughout its entire length. This secretion results principally from direct tactile or chemical stimulation of the mucosa by the chyme. Additional mucus is also secreted by the mucous cells in the intestinal pits, the crypts of Lieberkühn. This secretion is controlled mainly by local nervous reflexes.

SECRETION OF THE INTESTINAL DIGESTIVE JUICES—THE CRYPTS OF LIEBERKÜHN

Located on the entire surface of the small intestine are small pits called *crypts of Lieberkühn,* one of which is illustrated in Figure 64–13. The intestinal secretions are formed by the epithelial cells in these crypts at a rate of about 1800 ml per day. The secretions are almost pure extracellular fluid, and they have a slightly alkaline pH in the range of 7.5 to 8.0. They are rapidly reabsorbed by the villi. This circulation of fluid from the crypts to the villi obviously supplies a watery vehicle for absorption of substances from the chyme as it comes in contact with the villi, which is one of the primary functions of the small intestine.

Mechanism of Secretion of the Watery Fluid. The exact mechanism causing the marked secretion of watery fluid by the crypts of Lieberkühn is not known. However, it is believed that the secretion involves at least two active secretory processes, (1) active secretion of chloride ions into the crypts, and (2) active secretion of bicarbonate ions. The secretion of these ions, especially the chloride ions, causes electrical drag of sodium ions through the membrane as well. And, finally, all these ions together cause osmotic movement of water.

Effect of Cholera Toxin on Intestinal Secretion. Cholera toxin causes an extreme rate of fluid secretion into the crypts of Lieberkühn, especially in the jejunal region of the small intestine. During the first day of a cholera attack, as much as 15 liters of diarrhea fluid can be lost from the bowels, often leading to circulatory shock caused by dehydration and culminating in death within a few hours. Fortunately, though, most cases of cholera can be treated successfully simply by administering large amounts of saline and glucose solutions, either by mouth or intravenously.

The cholera toxin seems to have a specific effect in increasing active transport of chloride ions into the crypts of Lieberkühn; this in turn leads to the massive loss of fluid into the intestinal tract. Several other bacterial diseases of the intestines can also cause in-

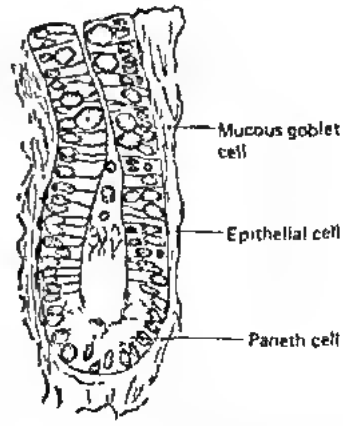


Figure 64–13. A crypt of Lieberkühn, found in all parts of the small intestine between the villi, which secretes almost pure extracellular fluid.

creased chloride transport similar to the effect of cholera, though usually they are less severe; these include infections caused by the colon bacillus, the dysentery bacillus, and others.

Enzymes in the Small Intestinal Secretion. When secretions of the small intestine are collected without cellular debris, they have almost no enzymes. However, the epithelial cells of the mucosa, especially those covering the villi, do contain digestive enzymes that are believed to digest food substances *while* they are being absorbed through the epithelium. These enzymes are the following: (1) several different *peptidases* for splitting small peptides into amino acids, (2) four enzymes for splitting disaccharides into monosaccharides—*sucrase, maltase, isomaltase,* and *lactase,* and (3) small amounts of *intestinal lipase* for splitting neutral fats into glycerol and fatty acids. Most, if not all, of these enzymes are mainly in the brush border of the epithelial cells. Therefore, they presumably catalyze hydrolysis of the foods on the outside surfaces of the microvilli prior to absorption of the end-products of digestion.

The epithelial cells deep in the crypts of Lieberkühn continually undergo mitosis, and the new cells gradually migrate along the basement membrane upward out of the crypts toward the tips of the villi where they are finally shed into the intestinal secretions. The life cycle of an intestinal epithelial cell is approximately 5 days. This rapid growth of new cells allows rapid repair of any excoriation that occurs in the mucosa.

REGULATION OF SMALL INTESTINAL SECRETION

Local Stimuli. By far the most important means for regulating small intestinal secretion are various local nervous reflexes, especially reflexes initiated by tactile or irritative stimuli. Therefore, for the most part, secretion in the small intestine occurs simply in response to the presence of chyme in the intestine—the greater the amount of chyme, the greater the secretion.

Hormonal Regulation. Some of the same hormones that promote secretion elsewhere in the gastrointestinal tract also increase small intestinal secretion, especially secretin and cholecystokinin. Also, some experiments suggest that other hormonal substances extracted from the small intestinal mucosa by the chyme might help to control secretion. However, in general, the local enteric reflex mechanisms probably play the dominant role.

SECRETIONS OF THE LARGE INTESTINE

Mucus Secretion. The mucosa of the large intestine, like that of the small intestine, is lined with crypts of Lieberkühn, but unlike the small intestine there are no villi. Also, the epithelial cells contain almost no enzymes. Instead, they are lined almost entirely by mucous cells that secrete only mucus. Also, on the surface epithelium of the large intestine are large numbers of mucous cells dispersed among the other epithelial cells.

Therefore, the great preponderance of secretion in the large intestine is mucus. This mucus contains large amounts of bicarbonate ions caused by active transport of these ions by the epithelial cells. The rate of secretion of mucus is regulated principally by direct, tactile stimulation of the mucous cells on the surface of the mucosa and by local nervous reflexes to the mucous cells in the crypts of Lieberkuhn. However, stimulation of the pelvic nerves, which carry the parasympathetic innervation to the distal one half to two thirds of the large intestine, also causes marked increase in the secretion of mucus. This occurs along with an increase in motility, which was discussed in the preceding chapter. Therefore, during extreme parasympathetic stimulation, often caused by emotional disturbances, so much mucus may be secreted into the large intestine that the person has a bowel movement of ropy mucus as often as every 30 minutes; the mucus contains little or no fecal material.

Mucus in the large intestine obviously protects the wall against excoriation, but, in addition, it provides the adherent medium for holding fecal matter together. Furthermore, it protects the intestinal wall from the great amount of bacterial activity that takes place inside the feces, and it, plus the alkalinity of the secretion (pH of 8.0 caused by large amounts of sodium bicarbonate), also provides a barrier to keep acids formed deep in the feces from attacking the intestinal wall.

Secretion of Water and Electrolytes in Response to Irritation. Whenever a segment of the large intestine becomes intensely irritated, as occurs when bacterial infection becomes rampant during *enteritis,* the mucosa then secretes large quantities of water and electrolytes in addition to the normal viscid solution of alkaline mucus. This acts to dilute the irritating factors and to cause rapid movement of the feces toward the anus. The usual result is *diarrhea* with loss of large quantities of water and electrolytes but also earlier recovery from the disease than would otherwise occur.

REFERENCES

Ackerman, S. H.. The developmental physiology of the stomach. Possible contributions to the regulatory disturbances of duodenal ulcer disease. *Proc Soc Exp Biol Med., 175* 119, 1984

Banks, P. A.. Pancreatitis. New York, Plenum Press, 1978.

Baron, J. H.. Clinical tests of Gastric Secretion. History, Methodology, and Interpretation. New York, Oxford University Press, 1978.

Berglindh, T.. The mammalian gastric parietal cell in vitro. *Annu Rev Physiol., 46* 377, 1984

Binder, H J (ed) Mechanisms of Intestinal Secretion. New York, A R. Liss, 1979
Bloom, S R. Gastrointestinal hormones *Int Rev Physiol*, *12* 71, 1977.
Bloom, S R (ed) Gut Hormones New York, Longman, Inc, 1978
Bonfils, S, *et al* Vagal control of gastric secretion *In* Crane, R K (ed) International Review of Physiology Gastrointestinal Physiology III Vol 19 Baltimore, University Park Press 1979, p 59
Brooks, F Diseases of the Exocrine Pancreas Philadelphia, W B. Saunders Co, 1980
Burland, W L., and Simkins, M A (eds) Cimetidine New York, Elsevier/North Holland 1977
Burnham, D B and Williams, J A. Stimulus-secretion coupling in pancreatic acinar cells. *J Pediatr Gastroenterol Nutr*, *3*(Suppl 1) S1, 1984
Davenport, H W Mechanisms of gastric and pancreatic secretion *In* Frohlich E D (ed) Pathophysiology, 2nd Ed Philadelphia, J B Lippincott Co, 1976 p 481
Flemstrom, G and Garner, A Soma characteristics of duodenal epithelium *Ciba Found Symp*, *109* 94, 1984
Forker, E L. Mechanisms of hepatic bile formation. *Annu Rev Physiol*, *39* 323 1977
Forte J G, *et al* Mechanisms of gastric H⁺ and Cl⁻ transport *Annu Rev Physiol*, *42* 111, 1980
Gardner, J Regulation of pancreatic exocrine function in vitro Initial steps in the actions of secretagogues *Annu Rev Physiol*, *41* 55, 1979.
Gerolomi, A and Sarles, J C Biliary secretion and motility *Int Rev Physiol*, *12* 223, 1977
Ginsberg, B L, and House, C R Stimulus-response coupling in gland cells *Annu Rev Biophys Bioeng* 9 55 1980
Grossman, M I Neural and hormonal regulation of gastrointestinal function An overview *Annu Rev Physiol*, *41* 27, 1979
Guth, P H Stomach blood flow and acid secretion *Annu Rev Physiol* *44* 3 1982
Hendrix, T R, and Paulk, H T Intestinal secretion *Int Rev Physiol*, *12* 257, 1977
Hersey B J, *et al* Cellular control of pepsinogen secretion *Annu Rev Physiol*, *46* 393 1984
Hirschowitz B I H-2 histamine receptors *Annu Rev Pharmacol Toxicol*, *19* 203, 1979
Jaffa, B M Hormones of the gastrointestinal tract *In* DeGroot, L J *et al* (eds.) Endocrinology Vol 3 New York, Grune & Stratton, 1979, p 1669
Jarzy Glass, G B Gastrointestinal Hormones New York, Raven Press, 1980
Johnson L. R. Gastrointestinal hormones and their functions *Annu Rev Physiol* *39* 135 1977
Jonve R S and Meyers, W C Regulation of hepatic biliary secretion *Annu Rev Physiol*, *41* 67, 1979
Lundgren O Microcirculation of the gastrointestinal tract and pancreas *In* Renkin, E. M, and Michel C C (eds) Handbook of Physiology Sec 2, Vol IV Bethesda, American Physiological Society 1984 p 799
Malinowska, D H, and Sachs, G Cellular mechanisms of acid secretion *Clin Gastroenterol*, *13* 309, 1984
Mason, D K Salivary Glands in Health and Disease Philadelphia W B Saunders Company, 1975
Miyoshi A (ed) Gut Peptides Secretion, Function, and Clinical Aspects New York, Elsevier/North Holland, 1979
Petersen, O H Electrophysiology of mammalian gland cells. *Physiol Rev*, *56* 535, 1976
Pucazo, J (ed) Glucagon in Gastroenterology Baltimore, University Park Press, 1978
Rehfeld, J F. Gastrointestinal hormones *In* Crane, R. K (ed) International Review of Physiology. Gastrointestinal Physiology III. Vol 19 Baltimore, University Park Press, 1979, p 291.
Rehfeld, J F, and Amdrup, E (eds) Gastrins and the Vagus New York, Academic Press, 1979
Reichen J, and Paumgartner, G Excretory function of the liver *In* Javitt, N B (ed) Internal Review of Physiology Liver and Biliary Tract Physiology I Vol 21 Baltimore, University Park Press, 1980, p 103
Rosselin, G, *et al* (eds) Hormone Receptors in Digestion and Nutrition. New York, Elsevier/North-Holland, 1979
Rothman, S S The digestive enzymes of the pancreas A mixture of inconstant proportions *Annu Rev Physiol*, *39* 373, 1977
Sachs, G, *et al* Gastric secretion *Int Rev Physiol*, *12* 127, 1977
Sachs, G *et al* H⁺ transport Regulation and mechanism in gastric mucosa and membrane vesicles *Physiol Rev*, *58* 106, 1978
Salen G, and Shefer, S Bile acid synthesis *Annu Rev Physiol*, *45* 679, 1983
Sarles H The exocrine pancreas. *Int Rev Physiol*, *12* 173, 1977.
Satir B The final steps in secretion *Sci Am* *233*(4) 28, 1975
Schulz, I and Blotzo, H H The exocrine pancreas The role of secretagogues, cyclic nucleotides and calcium in enzyme secretion *Annu Rev Physiol* *42* 127 1980
Shepherd A P Local control of intestinal oxygenation and blood flow *Annu Rev Physiol*, *44* 13, 1982
Soll, A, and Walsh, J H Regulation of gastric acid secretion *Annu Rev Physiol*, *41* 35, 1979
Strange R C Hepatic bile flow *Physiol Rev* *64* 1055, 1984
Stroud, R M., et al Mechanisms of zymogen activation *Annu Rev Biophys Bioeng*, *6* 177, 1977
Szurszewski J H Physiology of mammalian prevertebral ganglia *Annu Rev Physiol* *43* 53, 1981
Walsh, J H Circulating gastrin *Annu Rev Physiol*, *37* 81, 1975
Williams, J A. Regulatory mechanisms in pancreas and salivary acini *Annu Rev Physiol*, *46* 361, 1984
Willingham M C, and Pastan, I Endocytosis and exocytosis Current concepts of vesicle traffic in animal cells *Int Rev Cytol*, *92* 51, 1984
Young, J A Salivary secretion of inorganic electrolytes. *In* Crane, R K (ed) International Review of Physiology Gastrointestinal Physiology III Vol 19 Baltimore, University Park Press, 1979, p 1
Young, J A, and van Lennep, E W Morphology and physiology of salivary myoepithelial cells *Int Rev Physiol*, *12* 105, 1977

65

Digestion and Absorption in the Gastrointestinal Tract

LIBRARY

The foods on which the body lives, with the exception of small quantities of substances such as vitamins and minerals, can be classified as carbohydrates, fats, and proteins. However, these generally cannot be absorbed in their natural forms through the gastrointestinal mucosa and, for this reason, are useless as nutrients without the preliminary process of digestion Therefore, the present chapter discusses, first, the processes by which carbohydrates, fats, and proteins are digested into small enough compounds for absorption and, second, the mechanisms by which the digestive end-products, as well as water, electrolytes, and other substances, are absorbed.

DIGESTION OF THE VARIOUS FOODS

Hydrolysis as the Basic Process of Digestion. Almost all the carbohydrates of the diet are large *polysaccharides* or *disaccharides*, which are combinations of *monosaccharides* bound to each other by the process of *condensation*. This means that a hydrogen ion has been removed from one of the monosaccharides, while a hydroxyl ion has been removed from the next one; the two monosaccharides then are combined with each other at these sites of removal, and the hydrogen and hydroxyl ions combine to form water When the carbohydrates are digested back into monosaccharides, specific enzymes return the hydrogen and hydroxyl ions to the polysaccharides and thereby separate the monosaccharides from each other. This process, called *hydrolysis*, is the following:

$$R'' - R' + H_2O \xrightarrow[enzyme]{digestive} R''OH + R'H$$

Almost the entire fat portion of the diet consists of triglycerides (neutral fats), which are combinations of three *fatty acid* molecules condensed with a single *glycerol* molecule. In the process of condensation, three molecules of water had been removed. Digestion of the triglycerides consists of the reverse process, the fat-digesting enzymes returning molecules of water to the triglyceride molecule and thereby splitting the fatty acid molecules away from the glycerol. Here again, the process is one of *hydrolysis*

Finally, proteins are formed from *amino acids* that are bound together by *peptide linkages*. In this linkage a hydroxyl ion is removed from one amino acid, while a hydrogen ion is removed from the succeeding one; thus, the amino acids also combine together by a process of condensation, and digestion occurs by hydrolysis, the proteolytic enzymes returning the water to the protein molecules to split them into their constituent amino acids

Therefore, the chemistry of digestion is really simple, for in the case of all three major types of food, the same basic process of *hydrolysis* is involved. The only difference lies in the enzymes required to promote the reactions for each type of food.

All the digestive enzymes are proteins. Their secretion by the different gastrointestinal glands was discussed in the preceding chapter.

DIGESTION OF CARBOHYDRATES

The Carbohydrate Foods of the Diet. Only three major sources of carbohydrates exist in the normal human diet. These are sucrose, which is the disaccharide known popularly as cane sugar; lactose, which is a disaccharide in milk; and starches, which are large polysaccharides present in almost all nonanimal foods and particularly in the grains. Other carbohydrates ingested to a slight extent are glycogen, alcohol, lactic acid, pyruvic acid, pectins, dextrins, and minor quantities of other carbohydrate derivatives in meats. The diet also contains a large amount of cellulose, which is a carbohydrate. However, no enzymes capable of hydrolyzing cellulose are secreted by the human digestive tract. Consequently, cellulose cannot be considered to be a food for the human being, though it can be utilized by some lower animals.

Digestion of Carbohydrates in the Mouth. When food is chewed, it is mixed with the saliva, which contains the enzyme *ptyalin* (an α-amylase) secreted mainly by the parotid glands. This enzyme hydrolyzes starch into the disaccharide *maltose* and other small polymers of glucose containing 3 to 9 glucose molecules (such as *maltotriose* and a *limit dextrins* that are the branch points of the starch molecule), as shown in Figure

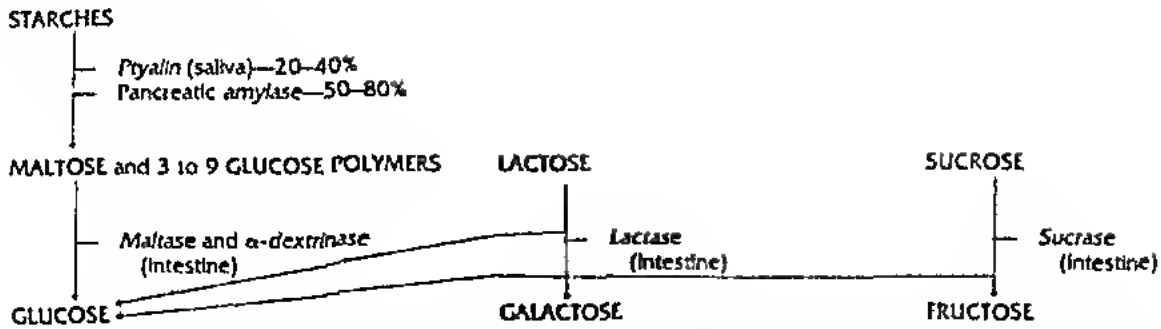


Figure 65–1. Digestion of carbohydrates.

65–1, but the food remains in the mouth only a short time, and probably not more than 3 to 5 per cent of all the starches that are eaten will have become hydrolyzed by the time the food is swallowed One can demonstrate the digestive action of ptyalin in the mouth by chewing a piece of bread for several minutes, after this time, the bread tastes sweet because of the maltose that has been liberated from the starches of the bread

Most starches in their natural state, unfortunately, are present in the food in small globules, each of which has a thin protective cellulose covering Therefore, most naturally occurring starches are digested only poorly by ptyalin unless the food is cooked to destroy the protective membrane.

Digestion of Carbohydrates in the Stomach. Even though food does not remain in the mouth long enough for ptyalin to complete the breakdown of starches into maltose, the action of ptyalin can continue for as long as an hour after the food has entered the stomach, that is, until the contents of the fundus are mixed with the stomach secretions Then the activity of the salivary amylase is blocked by the acid of the gastric secretions, for it is essentially nonactive as an enzyme once the pH of the medium falls below approximately 4 0 Nevertheless, on the average, before the food becomes completely mixed with the gastric secretions, as much as 30 to 40 per cent of the starches will have been hydrolyzed mainly to maltose

Digestion of Carbohydrates in the Small Intestine. ***Digestion by Pancreatic Amylase.*** Pancreatic secretion, like saliva, contains a large quantity of α-amylase which is almost identical in its function with the α-amylase of saliva and is capable of splitting starches into *maltose* and *other small glucose polymers.* Therefore, immediately after the chyme empties from the stomach into the duodenum and mixes with pancreatic juice, the starches that have not already been split are digested by amylase. In general, the starches are almost totally converted into maltose and other very small glucose polymers before they have passed beyond the jejunum.

Hydrolysis of Disaccharides and Small Glucose Polymers into Monosaccharides by the Intestinal Epithelial Enzymes. The epithelial cells lining the small intestine contain the four enzymes *lactase, sucrase, maltase,* and *α-dextrinase,* which are capable of splitting the disaccharides lactose, sucrose, and maltose, and the other small glucose polymers into their constituent monosaccharides These enzymes are located in the *brush border* of the cells lining the lumen of the intestine, and the disaccharides are digested as they come in contact with this border. The digested products, the monosaccharides, are then immediately absorbed into the portal blood Lactose splits into a molecule of *galactose* and a molecule of *glucose.* Sucrose splits into a molecule of *fructose* and a molecule of *glucose* Maltose and the other small glucose polymers all split into *molecules of glucose* Thus, the final products of carbohydrate digestion that are absorbed into the blood are all monosaccharides

In the ordinary diet, which contains far more starches than either sucrose or lactose, glucose represents more than 80 per cent of the final products of carbohydrate digestion, and galactose and fructose each rarely represent more than 10 per cent of the products of carbohydrate digestion

DIGESTION OF FATS

The Fats of the Diet. By far the most common fats of the diet are the neutral fats, also known as *triglycerides,* each molecule of which is composed of a glycerol nucleus and three fatty acids, as illustrated in Figure 65–2 Neutral fat is found in food of both animal origin and plant origin

In the usual diet are also small quantities of phospholipids, cholesterol, and cholesterol esters The phospholipids and cholesterol esters contain fatty acid, and, therefore, can be considered to be fats themselves. Cholesterol, on the other hand, is a sterol compound containing no fatty acid, but it does exhibit some of the physical and chemical characteristics of fats; it is derived from fats, and it is metabolized similarly to fats Therefore, cholesterol is considered from a dietary point of view to be a fat.

Digestion of Fats in the Intestine. A small amount of short chain triglycerides of butterfat origin is digested in the stomach by gastric lipase *(tributyrase)* However, the amount of digestion is so slight that it is unimportant Instead, essentially all fat digestion occurs in the small intestine as follows

Emulsification of Fat by Bile Acids. The first step in fat digestion is to break the fat globules into small sizes so that the water-soluble digestive enzymes can act on the globule surfaces This process is called emulsification of the fat, and it is achieved under the influence of *bile,* the secretion of the liver that does not contain any digestive enzymes. However, bile does contain a large quantity of *bile salts,* mainly in the form of ionized sodium salts, which are extremely important for the emulsification of fat. The carboxyl and other polar parts of the bile salt molecule are highly soluble in water, whereas most of the sterol portion of the bile salt is highly soluble in fat Therefore, the fat-soluble portion of the bile salt dissolves in the surface layer of the fat globule but with the polar portion of the salt projecting

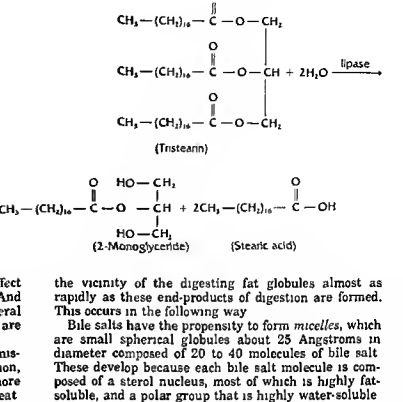


Figure 65–2. Hydrolysis of neutral fat catalyzed by lipase

outward and soluble in the surrounding fluids; this effect greatly decreases the interfacial tension of the fat And this property of the bile acids is multiplied even several more times by the small amounts of lecithin that are also in the bile.

When the interfacial tension of a globule of nonmiscible fluid is low, this nonmiscible fluid, on agitation, can be broken up into many minute particles far more easily than it can when the interfacial tension is great Consequently, a major function of the bile salts is to make the fat globules readily fragmentable by agitation in the small bowel. This action is the same as that of many detergents that are used widely in most household cleansers for removing grease.

Each time the diameters of the fat globules are decreased by a factor of 2 as a result of agitation in the small intestine, the total surface area of the fat increases two times. In other words, tha total surface area of tha fat particles in the intestinal contents is inversely proportional to the diameters of the particles Since the average size of the emulsified fat particles in the intestine is only 1 micron, this represents an increase of as much as 1000-fold in the total surface area of the fats caused by the emulsification process.

The lipases are water-soluble compounds and can attack the fat globules only on their surfaces Consequently, it can be readily understood how important this detergent function of bile salts is for the digestion of fats.

Digestion of Fats by Pancreatic Lipase. By far the most important enzyme for the digestion of fats is *pancreatic lipase* in the pancreatic juice. However, the epithelial cells of the small intestine also contain a minute quantity of lipase known as *enteric lipase*. Both of these act alike to cause hydrolysis of fat.

End-Products of Fat Digestion. Most of the triglycerides of the diet are split into *free fatty acids* and *2-monoglycerides*, as illustrated in Figure 65–3. However, small portions are not digested at all or remain in the diglyceride state.

Role of Bile Salts in Accelerating Fat Digestion—Formation of Micelles. The hydrolysis of triglycerides is a highly reversible process; therefore, accumulation of monoglycerides and free fatty acids in the vicinity of digesting fats very quickly blocks further digestion Fortunately, the bile salts play an important role in removing the monoglycerides and free fatty acids from the vicinity of the digesting fat globules almost as rapidly as these end-products of digestion are formed. This occurs in the following way

Bile salts have the propensity to form *micelles*, which are small spherical globules about 25 Angstroms in diameter composed of 20 to 40 molecules of bile salt These develop because each bile salt molecule is composed of a sterol nucleus, most of which is highly fat-soluble, and a polar group that is highly water-soluble The sterol nuclei of the 20 to 40 bile salt molecules of the micelle aggregate together to form a small fat globule in tha middle of the micelle This aggregation causes the polar groups to project outward to cover the surface of the micelle Since these polar groups are negatively charged, they allow the entire micelle globule to become dissolved in the water of the digestive fluids and to remain in stable solution despite the very large size of tha micelle

During triglyceride digestion, as rapidly as the monoglycerides and free fatty acids are formed they become dissolved in the fatty portion of the micelles, which immediately reduces these end-products of digestion in the vicinity of the digesting fat globules Consequently, the digestive process can proceed unabated

The bile salt micelles also act as a transport medium to carry the monoglycerides and the free fatty acids, both of which would otherwise be relatively insoluble, to the brush borders of the epithelial cells. There the monoglycerides and free fatty acids are absorbed, as will be discussed later On delivery of these substances to the brush border, the bile salts are again released back into the chyme to be used again and again for this "ferrying" process

Digestion of Cholesterol Esters and Phospholipids. Most of the cholesterol in the diet is in the form of cholesterol esters, which are combinations of free cholesterol and one molecule of fatty acid. And phospholipids also contain fatty acid chains within their molecules. Both the cholesterol esters and the phospholipids

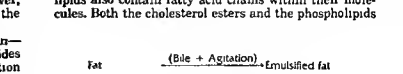


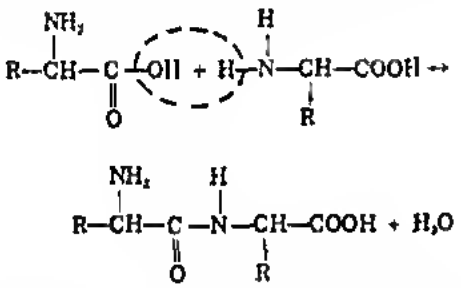
Figure 65–3. Digestion of fats.

are hydrolyzed by lipases in the pancreatic secretion that free the fatty acids—the enzyme *cholesterol ester hydrolase* to hydrolyze the cholesterol ester and *phospholipase A_2* to hydrolyze the phospholipid.

The bile salt micelles play identically the same role in "ferrying" free cholesterol as they play in "ferrying" monoglycerides and free fatty acids. Indeed, this role of the bile salt micelles is absolutely essential to the absorption of cholesterol because essentially no cholesterol is absorbed without the presence of bile salts. On the other hand, as much as 60 per cent of the triglycerides can be digested and absorbed even in the absence of bile salts.

DIGESTION OF PROTEINS

The Proteins of the Diet. The dietary proteins are derived almost entirely from meats and vegetables. These proteins in turn are formed of long chains of amino acids bound together by *peptide linkages*. A typical linkage is the following:



The characteristics of each type of protein are determined by the types of amino acids in the protein molecule and by the arrangement of these amino acids. The physical and chemical characteristics of the different proteins will be discussed in Chapter 69.

Digestion of Proteins in the Stomach. *Pepsin*, the important peptic enzyme of the stomach, is most active at a pH of 2 to 3 and is completely inactive at a pH above approximately 5. Consequently, for this enzyme to cause any digestive action on protein, the stomach juices must be acidic. It will be recalled from Chapter 64 that the gastric glands secrete a large quantity of hydrochloric acid. This hydrochloric acid is secreted by the oxyntic (parietal) cells at a pH of about 0.8, but, by the time it is mixed with the stomach contents and with the secretions from the nonoxyntic glandular cells of the stomach, the pH ranges around 2 to 3, a highly favorable range of acidity for pepsin activity.

Pepsin is capable of digesting essentially all the different types of proteins in the diet. One of the important features of pepsin digestion is its ability to digest collagen, an albuminoid that is affected little by other digestive enzymes. Collagen is a major constituent of the intercellular connective tissue of meats, and for the digestive enzymes of the digestive tract to penetrate meats and digest the cellular proteins it is first necessary that the collagen fibers be digested. Consequently, in persons lacking peptic activity in the stomach, the ingested meats are less well penetrated by the digestive enzymes and, therefore, may be poorly digested.

As illustrated in Figure 65–4, pepsin usually only begins the process of protein digestion, perhaps providing as much as 10 to 30 per cent of the total protein digestion. This splitting of proteins is a process of hydrolysis occurring at the peptide linkages between the amino acids.

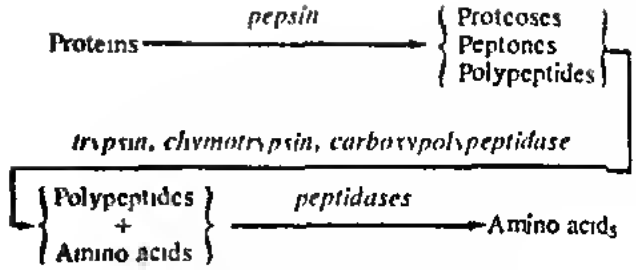


Figure 65–4. Digestion of proteins

Digestion of Proteins by Pancreatic Secretions. Most protein digestion occurs principally in the small intestine, under the influence of the proteolytic enzymes of the pancreatic secretion. When the proteins leave the stomach, they ordinarily are mainly in the form of proteoses, peptones, and large polypeptides. Immediately upon entering the small intestine, the partial breakdown products are attacked by the pancreatic enzymes *trypsin*, *chymotrypsin*, and *carboxypolypeptidase*, as illustrated in Figure 65–4. Both trypsin and chymotrypsin can split protein molecules into small polypeptides, carboxypolypeptidase then leaves individual amino acids from the carboxyl ends of the polypeptides.

Digestion of Peptides by the Epithelial Peptidases of the Small Intestine. The *brush border* of the small intestine contains several different enzymes for hydrolyzing the final peptide linkages of the remaining dipeptides and other small polypeptides as they come in contact with the epithelium of the villi. The enzymes responsible are *aminopolypeptidase* and several *dipeptidases*.

All the proteolytic enzymes—including those of the gastric juice, the pancreatic juice, and the brush border of the intestinal epithelial cells—are very specific for hydrolyzing individual types of peptide linkages. The linkages between certain pairs of amino acids differ in their bond energy and other physical characteristics from the linkages between other pairs. Therefore, a specific enzyme is required for specific types of linkages. *This accounts for the multiplicity of proteolytic enzymes.*

When food has been properly masticated and is not eaten in too large a quantity at any one time, about 98 per cent of all the proteins finally become either amino acids or dipeptides that can be absorbed into the blood.

BASIC PRINCIPLES OF GASTROINTESTINAL ABSORPTION

ANATOMICAL BASIS OF ABSORPTION

The total quantity of fluid that must be absorbed each day is equal to the ingested fluid (about 1.5 liters) plus that secreted in the various gastrointestinal secretions (about 7 liters). This comes to a total of approximately 8 to 9 liters. All but 1.5 liters of this is absorbed in the small intestine, leaving only 1.5 liters to pass through the ileocecal valve into the colon each day.

The stomach is a poor absorptive area of the gastrointestinal tract because it lacks the typical

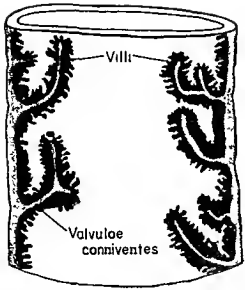


Figure 65–5. A longitudinal section of the small intestine, showing the valvulae conniventes covered by villi

villus type of absorptive membrane and also because the junctions between the epithelial cells are tight junctions. Only a few highly lipid-soluble substances, such as alcohol and some drugs like aspirin, can be absorbed in small quantities.

The Absorptive Surface of the Intestinal Mucosa—The Villi. Figure 65–5 illustrates the absorptive surface of the intestinal mucosa, showing many folds called *valvulae conniventes* (or *folds of Kerckring*), which increase the surface area of the absorptive mucosa about threefold. These folds extend circularly most of the way around the intestine and are especially well developed in the duodenum and jejunum, where they often protrude as much as 8 mm into the lumen.

Located over the entire surface of the small intestine, from approximately the point at which the common bile duct empties into the duodenum down to the ileocecal valve, are literally millions of small *villi*, which project about 1 mm from the surface of the mucosa, as shown on the surfaces of the valvulae conniventes in Figure 65–5 and in detail in Figure 65–6. These villi lie so close to each other in the upper small intestine that they actually touch in most areas, but their distribution is less profuse in the distal small intestine. The presence of villi on the mucosal surface enhances the absorptive area another tenfold.

The intestinal epithelial cells are characterized by a brush border, consisting of about 600 *microvilli* 1 μm in length and 0.1 μm in diameter protruding from each cell; these are illustrated in the electron micrograph in Figure 65–7. This increases the surface area exposed to the intestinal materials another 20-fold. Thus, the combination of the folds of Kerckring, the villi, and the microvilli increases the absorptive area of the mucosa about 600-fold, making a tremendous total area of about 250 square meters for the entire small intestine—about the surface area of a tennis court.

Figure 65–6A illustrates the general organization of a villus, emphasizing especially the advantageous arrangement of the vascular system for absorption of fluid and dissolved material into the portal blood, and the arrangement of the *central lacteal* for absorption into the lymph. Figure 65–6B shows the cross-section of a villus, and Figure 65–7 shows many small *pinocytic* vesicles, which are pinched-off portions of infolded epithelium surrounding extracellular materials that have been entrapped inside the cells. Small amounts of substances are absorbed by this physical process of *pinocytosis*, though, as noted later in the chapter, most absorption occurs by means of single molecular transfer. Located near the brush border of the epithelial cell are many *mitochondria*, which supply the cell with oxidative energy needed for *active transport* of materials through the intestinal epithelium; this also is discussed later in the chapter. Also, extending linearly into each microvillus of the brush border

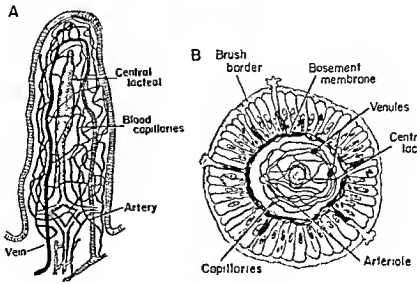


Figure 65–6. Functional organization of the villus. A, Longitudinal section. B, Cross-section showing the epithelial cells and basement membrane.

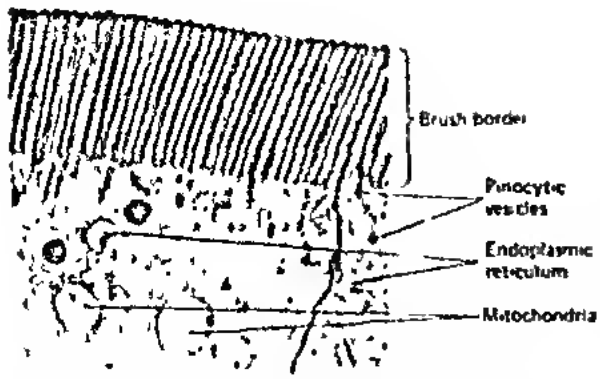


Figure 65–7. Brush border of the gastrointestinal epithelial cell, showing, also, pinocytic vesicles, mitochondria, and endoplasmic reticulum lying immediately beneath the brush border. (Courtesy of Dr. William Lockwood.)

are multiple actin filaments that are believed to contract and cause continual movement of the microvilli, keeping them constantly exposed to new quantities of intestinal fluid.

BASIC MECHANISMS OF ABSORPTION

Absorption through the gastrointestinal mucosa occurs by *active transport* and by *diffusion*, as is also true for other membranes. The physical principles of these processes were explained in Chapter 4.

Briefly, active transport imparts energy to the substance as it is being transported for the purpose of concentrating it on the other side of the membrane or for moving it against an electrical potential. On the other hand, the term diffusion means simply transport of substances through the membrane as a result of molecular movement *along*, rather than against, an electrochemical gradient.

ABSORPTION IN THE SMALL INTESTINE

Normally, absorption from the small intestine each day consists of several hundred grams of carbohydrates, 100 or more grams of fat, 50 to 100 grams of amino acids, 50 to 100 grams of ions, and 7 to 8 liters of water. However, the absorptive *capacity* of the small intestine is far greater than this: as much as several kilograms of carbohydrates per day, 500 to 1000 grams of fat per day, 500 to 700 grams of amino acids per day, and 20 or more liters of water per day. In addition, the large intestine can absorb still more water and ions, though almost no nutrients.

ABSORPTION OF WATER

Isosmotic Absorption. Water is transported through the intestinal membrane entirely by the process of *diffusion*. Furthermore, this diffusion obeys the usual laws of osmosis. Therefore, when the chyme is dilute, water is absorbed through the intestinal mucosa into the blood of the villi by osmosis.

On the other hand, water can also be transported in the opposite direction, from the plasma into the chyme. This occurs especially when hyperosmotic solutions are discharged from the stomach into the duodenum. Usually within minutes, sufficient water is transferred by osmosis to make the chyme isosmotic with the plasma. Thereafter, the chyme remains almost exactly isosmotic throughout its total passage through the small and large intestines.

As dissolved substances are absorbed from the lumen of the gut into the blood, the absorption tends to decrease the osmotic pressure of the chyme, but water diffuses so readily through the intestinal membrane (because of large 7 to 15 Å intercellular pores through the so-called "tight junctions" between the epithelial cells) that it almost instantaneously "follows" the absorbed substances into the blood. Therefore, as ions and nutrients are absorbed, so also is an isosmotic equivalent of water absorbed. In this way not only are the ions and nutrients almost entirely absorbed before the chyme passes through the intestinal tract, but so also is almost 99 per cent of the water absorbed.

ABSORPTION OF IONS

Active Transport of Sodium. Twenty to 30 grams of sodium are secreted into the intestinal secretions each day. In addition, the normal person eats 5 to 8 grams of sodium each day. Combining these two, the small intestine absorbs 25 to 35 grams of sodium each day, which amounts to about one seventh of all the sodium that is present in the body. One can well understand that whenever the intestinal secretions are lost to the exterior, as in extreme diarrhea, the sodium reserves of the body can be depleted to a lethal level within hours. Normally, this sodium is secreted and reabsorbed continually with only about 1 milliequivalent lost in the feces each day. The sodium plays an important role in the absorption of sugars and amino acids, as we shall see in subsequent discussions.

The basic mechanism of sodium absorption from the intestine is illustrated in Figure 65–8. The principles of this mechanism, which were discussed in Chapter 9, are also essentially the same as those for absorption of sodium from the renal tubules, as discussed in Chapter 34. The motive power for the sodium absorption is provided by active transport of sodium from inside the epithelial cells through the side walls of these cells into the intercellular spaces. This is illustrated by the heavy black arrows in Figure 65–8. This active transport obeys the usual laws of active transport: it requires energy, and it is catalyzed by appropri-

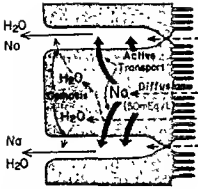


Figure 65–8. Absorption of sodium through the intestinal epithelium. Note also the osmotic absorption of water—that is, the water "follows" the sodium through the epithelial membrane.

ate ATPase carrier enzymes in the cell membrane (see Chapter 9). Part of the sodium is transported along with chloride ions that are passively "dragged" along by the positive electrical charges of the sodium ion. However, other sodium ions are absorbed while either potassium or hydrogen ions are transported into the lumen of the gut in exchange for the sodium ions. In the membrane of the brush border are special transport proteins that facilitate these exchanges between sodium and potassium or sodium and hydrogen

The active transport of sodium reduces its concentration in the cell to a low value (about 50 mEq/liter), as also illustrated in Figure 65–8. Since the sodium concentration in the chyme is normally about 142 mEq/liter (that is, approximately equal to that in the plasma), sodium moves by passive absorption from the chyme through the brush border of the epithelial cell into the epithelial cell cytoplasm. This replaces the sodium that is actively transported out of the epithelial cells into the intercellular spaces.

The next step in the transport process is osmosis of water into the intercellular spaces. This movement is caused by the osmotic gradient created by the elevated concentration of ions in the intercellular space. Most of this osmosis occurs through the "tight junctions" between the apical borders of the epithelial cells, as discussed earlier, but a smaller proportion occurs through the cells themselves. The osmotic movement of water creates a flow of fluid into the intercellular space, then through the basement membrane of the epithelium, and finally into the circulating blood of the villi.

Absorption of Chloride Ions in the Duodenum and Jejunum. In the upper part of the small intestine chloride absorption is mainly by passive diffusion. The absorption of sodium ions through the epithelium creates electronegativity in the chyme and electropositivity on the basal side of the epithelial cells. Then chloride ions move along this electrical gradient to "follow" the sodium ions.

"Active" Absorption of Bicarbonate Ions in the Duodenum and Jejunum. Often, large quantities of bicarbonate ions must be reabsorbed from the upper small intestine because of the large amounts of bicarbonate ions in both the pancreatic secretion and bile. However, the bicarbonate ion is absorbed in an indirect way as follows: When sodium ions are absorbed, moderate amounts of hydrogen ions are secreted into the lumen of the gut in exchange for some of the sodium, as explained earlier. These hydrogen ions in turn combine with the bicarbonate ion to form carbonic acid (H_2CO_3), and this then dissociates to form H_2O and CO_2. The water remains part of the chyme in the intestines, but the carbon dioxide is readily absorbed into the blood and subsequently expired through the lungs. Thus, this is the so-called "active" absorption of bicarbonate ions. It is the same mechanism that occurs in the tubules of the kidneys.

Active Absorption of Chloride Ions and Active Secretion of Bicarbonate Ions in the Ileum and Large Intestine. The epithelial cells of the ileum and of the large intestine have the special capability of actively absorbing chloride ions by means of a tightly coupled transport mechanism in which an equivalent number of bicarbonate ions are secreted. The functional role of this mechanism is to provide bicarbonate ions for neutralization of acidic products formed by bacteria—especially in the large intestine.

Various bacterial toxins, particularly those of cholera, colon bacilli, and staphylococci, can strongly stimulate this chloride-bicarbonate exchange mechanism.

Absorption of Other Ions. Calcium ions are actively absorbed, especially from the duodenum, and calcium ion absorption is exactly controlled in relation to the need of the body for calcium. One important factor controlling calcium absorption is parathyroid hormone secreted by the parathyroid glands, and another is vitamin D. The parathyroid hormone activates vitamin D in the kidneys, and the activated vitamin D in turn greatly enhances calcium absorption. These effects are discussed in Chapter 79.

Iron ions are also actively absorbed from the small intestine. The principles of iron absorption and the regulation of its absorption in proportion to the body's need for iron were discussed in Chapter 4.

Potassium, magnesium, phosphate, and probably still other ions can also be actively absorbed through the mucosa. In general, the monovalent ions are absorbed with ease and in great quantities. On the other hand, the bivalent ions are normally absorbed in only small amounts; for instance, the maximum absorption of calcium ions is only 1/50 as great as the normal absorption of sodium ions. Fortunately, only small quantities of the divalent ions are normally needed by the body.

ABSORPTION OF NUTRIENTS

Absorption of Carbohydrates

Essentially all the carbohydrates are absorbed in the form of monosaccharides, only a small fraction of a per cent being absorbed as disaccharides and almost none as larger carbohydrote compounds. Furthermore, little carbohydrate absorption results from simple diffusion, for the pores of the mucosa through which diffusion occurs are essentially impermeable to water-soluble solutes with molecular weights greater than 100.

That the transport of most monosaccharides through the intestinal membrane is an active process is demonstrated by several important experimental observations:

1. Transport of most of them, especially glucose and galactose, can be blocked by metabolic inhibitors, such as iodoacetic acid, cyanides, and phlorhizin

2 The transport is selective, specifically transporting certain monosaccharides without transporting others. The order of preferenca for transporting different monosaccharides and their relative rates of transport in comparison with glucosa are:

Galactose	1.1
Glucose	1.0
Fructosa	0.4
Mannose	0.2
Xylose	0.15
Arabinose	0.1

3 There is a maximum rate of transport for each type of monosaccharide. The most rapidly transported monosaccharide is galactose, with glucose running a close second Fructose, which is also one of tha three important monosaccharides for nutrition, is absorbed less than half as rapidly as either galactose or glucose; also, its mechanism of absorption is different, as will be explained below.

4 There is competition between certain sugars for the respective carrier system. For instance, if large amounta of galactose are being transported, the amount of glucose that can be transported simultaneously is considerably reduced.

Mechanism of Glucose and Galaetose Absorption. Glurose and galactose transport either ceases or is greatly reduced wherever active sodium transport is blocked. Therefore, it is assumed that the energy required for transport of these two monosaccharides is actually provided by the sodium transport system. A theory that attempts to explain this is the following: It is known that the carrier protein for transport of glucose (which is the carrier for galactose as well) is present in the brush border of the epithelial cell. However, this carrier will not transport the glucose in the absence of sodium transport. Therefore, it is believed that the carrier protein has receptor sites for both a glucose molecule and a sodium ion, and that it will not transport either of these to the interior of the epithelial cell until both receptor sites are simultaneously filled. The energy to cause movement of the carrier from the exterior of the membrane to the interior is derived from the difference in sodium concentration between the outside and inside. That is, as sodium diffuses to the inside of the cell it "drags" the glucose along with it, thus providing the energy for transport of the glucose. For obvious reasons, this explanation is called the *sodium cotransport theory* for glurose transport; it is also called *secondary active transport* of glucose. This sodium cotransport of glucose obviously moves the glucose only to the interior of the cell. However, this increases the intracellular glucose concentration to a higher than normal level, and the glucose then diffuses, probably by facilitated diffusion, through the basolateral membrane of the epithelial cell into the extracellular fluid.

Subsequently, we will see that sodium transport is also required for transport of many if not all amino acids, suggesting a similar "carrier-drag" mechanism for amino acid transport.

Absorption of Fructose. Transport of fructose is slightly different from that of most other monosaccharides. It is not blocked by some of the same metabolic poisons—specifically, phlorhizin—and it does not require metabolic energy for transport, even though it does require a specific carrier. Therefore, it is transported by *facilitated diffusion* rather than active transport. Also, it is mainly converted into glucose inside the epithelial cell before entering the portal blood, the fructose first becoming phosphorylated, then converted to glucose, and finally released from the epithelial cell into the blood.

Absorption of Proteins

Most proteins are absorbed in the form of amino acids. However, small quantities of dipeptides and even tripeptides are also absorbed, and extremely minute quantities of whole proteins can at times be absorbed by the process of pinocytosis, though not by the usual absorptive mechanisms.

The absorption of amino acids also obeys the principles listed above for active absorption of glucose; that is, the different types of amino acids are absorbed selectively and certain ones interfere with the absorption of others, illustrating that common carrier systems exist. Finally, metabolic poisons block the absorption of amino acids in the same way that they block the absorption of glucose.

Absorption of amino acids through the intestinal mucosa can occur far more rapidly than can protein digestion in the lumen of the intestine. As a result, the normal rate of absorption is determined not by the rate at which they can be absorbed but by the rate at which they can be released from the proteins during digestion. For these reasons, essen-

tially no free amino acids can be found in the intestine during digestion—that is, they are absorbed as rapidly as they are formed. Since most protein digestion occurs in the upper small intestine, most protein absorption occurs in the duodenum and jejunum.

Basic Mechanisms of Amino Acid Transport. As is true for monosaccharide absorption, very little is known about the basic mechanisms of amino acid transport. However, at least four different carrier systems transport different amino acids—one transports *neutral amino acids*, a second transports *basic amino acids*, a third transports *acidic amino acids*, and a fourth has specificity for the two imino acids *proline* and *hydroxyproline*. Also, the transport mechanisms have far greater affinity for transporting L-stereoisomers of amino acids than D-stereoisomers.

Amino acid transport (at least for most of the amino acids), like glucose transport, occurs only in the presence of simultaneous sodium transport. Furthermore, the carrier systems for amino acid transport, like those for glucose transport, are in the brush border of the epithelial cell. It is believed that amino acids are transported by the same *sodium cotransport* mechanism as that explained above for glucose transport. That is, the theory postulates that the carrier has receptor sites for both an amino acid molecule and a sodium ion. Only when both of the sites are filled will the carrier move both the sodium and the amino acid to the interior of the cell at the same time. Because of the sodium gradient across the brush border, the sodium diffusion to the cell interior pulls the amino acid to the interior where the amino acid becomes trapped. Therefore, amino acid concentration increases within the cell, and it then diffuses through the sides or base of the cell into the portal blood, probably by a facilitated diffusion process

Absorption of Fats

Earlier in this chapter it was pointed out that as fats are digested to form monoglycerides and free fatty acids, both of these digestive end-products become dissolved in the lipid portion of the bile acid micelles. Because of the molecular dimensions of these micelles, only 2.5 nanometers in diameter, and also because of their highly charged exterior, they are soluble in the chyme. In this form the monoglycerides and the fatty acids are transported to the surfaces of the brush border microvilli, even penetrating into the recesses among the moving, agitating microvilli. On coming in contact with these surfaces, both the monoglycerides and the fatty acids immediately diffuse through the epithelial membrane, because they are equally as soluble in this membrane as in the micelles. This leaves the bile acid micelles still in the chyme. The micelles then diffuse back through the chyme and absorb still more monoglycerides and fatty acids, and similarly transport these also to the epithelial cells. Thus, the bile acids perform a "ferrying" function, which is highly important for fat absorption. In the presence of an abundance of bile acids, approximately 97 per cent of the fat is absorbed; in the absence of bile acids, only 50 to 60 per cent is normally absorbed.

The mechanism for absorption of the monoglycerides and fatty acids through the brush border is based entirely on the fact that both these substances are highly lipid-soluble. Therefore, they become dissolved in the membrane and simply diffuse to the interior of the cell.

The undigested triglycerides and the diglycerides are both also highly soluble in the lipid membrane of the epithelial cell. However, only small quantities of these are normally absorbed because the bile acid micelles will not dissolve either triglycerides or diglycerides and therefore will not ferry them to the epithelial membrane.

After entering the epithelial cell, the fatty acids and monoglycerides are taken up by the smooth endoplasmic reticulum, and here they are mainly recombined to form new triglycerides. However, a few of the monoglycerides are further digested into glycerol and fatty acids by an epithelial cell lipase. Then, the free fatty acids are reconstituted by the smooth endoplasmic reticulum into triglycerides. Most of the glycerol that is utilized for this purpose is synthesized *de novo* from alpha-glycerophosphate, this synthesis requiring both energy from ATP and a complex of enzymes to catalyze the reactions.

Once formed, the triglycerides aggregate within the endoplasmic reticulum into globules along with absorbed cholesterol, absorbed phospholipids, and small amounts of newly synthesized cholesterol and phospholipids. The phospholipids arrange themselves in these globules with the fatty portion of the phospholipid toward the center and the polar portions located on the surface. This provides an electrically charged surface that makes these globules miscible with the fluids of the cell. In addition, small amounts of β-*lipoprotein*, also synthesized by the endoplasmic reticulum, coat part of the surface of each globule. In this form the globule diffuses to the side of the epithelial cell and is excreted by the process of cellular *exocytosis* into the space between the cells; from there it passes into the lymph in the central lacteal of the villus. These globules are then called *chylomicrons*.

The β-lipoprotein is essential for cellular exocytosis of the chylomicrons to occur, because this protein provides a means for attaching the fatty globule to the cell membrane before it is extruded. In persons who have a genetic inability to form this β-lipoprotein, the epithelial cells become engorged with fatty products that cannot proceed the rest of the way to be absorbed.

Transport of the Chylomicrons in the Lymph. From the sides of the epithelial cells the chylomicrons wend their way into the central lacteals of the villi and from here are propelled, along with the lymph, by the lymphatic pump upward through the thoracic duct to be emptied into the great veins of the neck. Between 80 and 90 per cent of all fat absorbed from the gut is absorbed in this manner and is transported to the blood by way of the thoracic lymph in the form of chylomicrons.

Direct Absorption of Fatty Acids into the Portal Blood. Small quantities of short chain fatty acids, such as those from butterfat, are absorbed directly into the portal blood rather than being converted into triglycerides and absorbed into the lymphatics The cause of this difference between short and long chain fatty acid absorption is that the shorter chain fatty acids are more water-soluble and are not reconverted into triglycerides by the endoplasmic reticulum. This allows direct diffusion of these fatty acids from the epithelial cells into the capillary blood of the villus

ABSORPTION IN THE LARGE INTESTINE: FORMATION OF THE FECES

Approximately 1500 ml of chyme pass through the ileocecal valve into the large intestine each day Most of the water and electrolytes in this are absorbed in the colon, usually leaving less than 100 ml of fluid to be excreted in the feces Also, essentially all the ions are also absorbed, leaving only about 1 mEq each of sodium and chloride ions to be lost in the feces.

Most of the absorption in the large intestine occurs in the proximal half of the colon, giving this portion the name *absorbing colon*, whereas the distal colon functions principally for storage and is therefore called the *storage colon*

Absorption and Secretion of Electrolytes and Water. The mucosa of the large intestine, like that of the small intestine, has a high capability for active absorption of sodium, and the electrical potential created by the absorption of the sodium causes chloride absorption as well The "tight junctions" between the epithelial cells of the large intestinal epithelium are much tighter than those of the small intestine. This prevents significant amounts of back-diffusion of ions through these junctions, thus allowing the large intestinal mucosa to absorb sodium ions far more completely—that is, against a much higher concentration gradient—than can occur in the small intestine.

In addition, as in the distal portion of the small intestine, the mucosa of the large intestine actively secretes *bicarbonate ions* while it simultaneously actively absorbs an equal amount of chloride ions in an exchange transport process. The bicarbonate helps neutralize the acidic end-products of bacterial action in the colon.

The absorption of sodium and chloride ions creates an osmotic gradient across the large intestinal mucosa, which in turn causes absorption of water.

Bacterial Action In the Colon. Numerous bacteria, especially colon bacilli, are present in the absorbing colon. These are capable of digesting small amounts of cellulose, in this way providing a few calories of nutrition to the body each day. In herbivorous animals this source of energy is very significant, though it is of negligible importance in the human being Other substances formed as a result of bacterial activity are vitamin K, vitamin B_{12}, thiamin, riboflavin, and various gases that contribute to *flatus* in the colon—especially carbon dioxide, hydrogen gas, and methane. Vitamin K is especially important, for the amount of this vitamin in the ingested foods is normally insufficient to maintain adequate blood coagulation.

Composition of the Feces. The feces normally are about three-fourths water and one-fourth solid matter composed of about 30 per cent dead bacteria, 10 to 20 per cent fat, 10 to 20 per cent inorganic matter, 2 to 3 per cent protein, and 30 per cent undigested roughage of the food and dried constituents of digestive juices, such as bile pigment and sloughed epithelial cells. The large amount of fat derives mainly from fat formed by bacteria and fat in the sloughed epithelial cells.

The brown color of feces is caused by *stercobilin* and *urobilin*, which are derivatives of bilirubin The odor is caused principally by the products of bacterial action; these vary from one person to another, depending on each person's colonic bacterial flora and on the type of food eaten The actual odoriferous products include *indole, skatole, mercoptans*, and *hydrogen sulfide*

REFERENCES

Anderson, P O Vascular control in the colon and rectum. *Scand J Gastroenterol* 21 65 1984

Baker P F and Knight, D E Chemiosmotic hypotheses of exocytosis: A critique *Biosci Rep*, *4* 285, 1984

Bickel, H (ed.) Digestion and Absorption of Nutrients. Ft Lee, NJ, J K Burgess, 1983

Cane, R N The role of Ca^{2+} stores in secretion *Cell-Calcium*, *5* 89, 1984

Christensen, H N The regulation of amino acid and sugar absorption by diet *Nutr Rev*, *42* 237, 1984

Creamer, B Intestinal structure in relation to absorption *Biomembranes*, *4A* 1, 1974

Cummings, J H Colonic absorption The importance of short chain fatty acids in man *Scand J Gastroenterol (Suppl)*, *93* 89, 1984

Davenport, H W A Digest of Digestion, 2nd Ed. Chicago, Year Book Medical Publishers, 1978

Donowitz, M, *et al* Cytosol free Ca^{++} in the regulation of active intestinal Na and Cl transport. KROC Found Ser., *17* 171, 1984

Flemstrom, G., and Garner, A Some characteristics of duodenal epithelium *Ciba Found. Symp*, *109* 94, 1984

Forth, W, and Rummel, W. Iron absorption. *Physiol Rev*, *53* 724 1973

Frizzell, R. A, and Schultz, S G Models of electrolyte absorption and secretion by gastrointestinal epithelia *In* Crane, R. K (ed.) International Review of Physiology Gastrointestinal Physiology III. Vol 19 Baltimore, University Park Press, 1979, p 205

Gardner, M L. Intestinal assimilation of intact peptides and proteins from the diet—a neglected field? *Biol Rev*, *59* 289, 1984

Holzer, H., and Tschesche, H (eds.) Biological Functions of Proteinases. New York, Springer-Verlag, 1979

Jackson, M J., Treatment of short chain fatty acids *Biomembranes*, *4B* 673, 1974

Kenny, A. J., and Maroux, S. Topology of microvillar membrane hydrolases of kidney and intestine *Physiol. Rev*, *62* 91, 1982
Kim, Y S. *et al*: Intestinal peptide hydrolases. Peptide and amino acid absorption *Med Clin. North Am*, *58* 1397, 1974
Kotyk, A · Mechanisms of nonelectrolyte transport. *Biochim Biophys Acta*, *300* 183, 1973
Levitan, R., and Wilson, D E.. Absorption of water soluble substances. *In* MPT International Review of Physiology Vol 4. Baltimore, University Park Press, 1974, p 293
Mailman, D. Relationships between intestinal absorption and hemodynamics. *Annu. Rev. Physiol*, *44* 43, 1982
Matthews, D M.. Intestinal absorption of amino acids and peptides. *Proc. Nutr Soc*. *31*,171, 1972
Matthews, D M · Absorption of amino acids and peptides from the intestine. *Clin Endocrinol. Metabol*. *3* 3, 1974
Matthews, D M: Absorption of water soluble vitamins *Biomembranes*, *4B* 847, 1974
Matthews, D M: Intestinal absorption of peptides. *Physiol Rev*, *55*.537, 1975
Norum, K. R, *et al* Transport of cholesterol. *Physiol Rev*, *63* 1343, 1983
Ockner, R. K, and Isselbacher, K J Recent concepts of intestinal fat absorption *Rev Physiol Biochem. Pharmacol*, *71* 107, 1984
Okuda, K.: Intestinal mucosa and vitamin B_{12} absorption *Digestion*, *6*.173, 1972.
Olsen, W. A.. Carbohydrate absorption. *Med Clin North Am*, *58* 1387, 1974
Rose, R C Water-soluble vitamin absorption in intestine *Annu Rev Physiol*, *42*:157, 1980
Schultz, S. G · A cellular model for active sodium absorption by mammalian colon. *Annu Rev Physiol*, *46* 435, 1984
Schultz, S G.. Principles of electrophysiology and their application to epithelial tissues. *In* MPT International Review of Science Physiology. Vol 4, Baltimore, University Park Press, 1974, p 69
Schultz, S G, *et al* Ion transport by mammalian small intestine *Annu Rev Physiol*, *35* 51, 1974
Silk, D B A, and Dawson, A. M Intestinal absorption of carbohydrate and protein in man *In* Crane, R K (ed) International Review of Physiology: Gastrointestinal Physiology III Vol 19 Baltimore, University Park Press, 1979, p 151
Simmonds, W J Absorption of lipids. *In* MPT International Review of Science Physiology. Vol 4 Baltimore, University Park Press, 1974, p 343
Smith, P L., and McCabe, R. D Mechanisms and regulation of transcellular potassium transport by the colon *Am. J Physiol*., *247* G445, 1984
Smyth, D H. (ed) Intestinal Absorption. Vols 4A and 4B New York, Plenum Press, 1974
Soergel, K H., and Hofmann, A F Absorption *In* Frohlich, E D (ed) Pathophysiology, 2nd Ed Philadelphia, J B Lippincott Co, 1976, p 499
Stevens, B R., *et al* Intestinal transport of amino acids and sugars Advances using membrane vesicles. *Annu Rev Physiol* *46* 417, 1984
Turnberg, L. A Absorption and secretion of salt and water by the small intestine *Digestion*, *9* 357, 1973
Ugolev, A M Membrane (contact) digestion. *Biomembranes*, *4A* 285, 1974
Van Campen, D Regulation of iron absorption *Fed Proc* *33* 100, 1974
Watson, D W, and Sodeman, W A, *Jr* The small intestine *In* Sodeman, W A *Jr* and Sodeman, T M (eds.) Pathologic Physiology Mechanisms of Disease 6th Ed Philadelphia, W B Saunders Company, 1979, p 824
Wiseman, G Absorption of protein digestion products *Biomembranes* *4A* 363, 1974

66

Physiology of Gastrointestinal Disorders

The logical treatment of most gastrointestinal disorders depends on a basic knowledge of gastrointestinal physiology. The purpose of this chapter, therefore, is to discuss a few representative types of malfunction that have special physiological bases or consequences.

DISORDERS OF SWALLOWING AND OF THE ESOPHAGUS

Paralysis of the Swallowing Mechanism. Damage to the 5th, 9th, or 10th nerve can cause paralysis of significant portions of the swallowing mechanism. Also, a few diseases, such as poliomyelitis or encephalitis, can prevent normal swallowing by damaging the swallowing center in the brain stem. Finally, malfunction of the swallowing muscles, as occurs in *muscle dystrophy* or in failure of neuromuscular transmission in *myasthenia gravis* or *botulism*, can also prevent normal swallowing.

When the swallowing mechanism is partially or totally paralyzed, the abnormalities that can occur include (1) complete abrogation of the swallowing act so that swallowing cannot occur at all, (2) failure of the glottis to close so that food passes into the lungs instead of the esophagus, or (3) failure of the soft palate and uvula to close the posterior nares so that food refluxes into the nose during swallowing.

One of the most serious instances of paralysis of the swallowing mechanism occurs when patients are under deep anesthesia. Often they vomit large quantities of materials from the stomach into the pharynx; then, instead of swallowing the materials again, they simply suck them into the trachea because the anesthetic has blocked the reflex mechanism of swallowing. As a result, such patients occasionally choke to death on their own vomitus.

Achalasia and Megaesophagus. Achalasia is a condition in which the lower esophageal sphincter fails to relax during the swallowing mechanism. As a result, food transmission from the esophagus into the stomach is impeded or prevented. Pathological studies have shown the physiological basis of this condition to be either pathology of or absence of the myenteric plexus in the lower two thirds of the esophagus. The musculature of the lower esophagus instead remains incoordinately contracted, and the myenteric plexus has lost the ability to transmit a signal to cause "receptive relaxation" of the gastroesophageal sphincter as food approaches this area during the swallowing process.

When achalasia becomes severe, the esophagus may not empty the swallowed food into the stomach for many hours, instead of within a few seconds which is the normal time. Over months and years, the esophagus becomes tremendously enlarged until it often can hold as much as one liter of food, which becomes putridly infected during the long periods of esophageal stasis. The infection may also cause ulceration of the esophageal mucosa, sometimes leading to severe substernal pain or even rupture and death. The food often refluxes into the pharynx and sometimes is then aspirated into the lungs, causing aspiration pneumonia. Fortunately, considerable benefit can be achieved by stretching the lower end of the esophagus by means of a balloon inflated on the end of a swallowed esophageal tube. Also, antispasmotic drugs (drugs that relax smooth muscle), such as hydralazine and nifedipine, can also be helpful.

DISORDERS OF THE STOMACH

Gastritis. Gastritis means inflammation of the gastric mucosa. This is exceedingly common in the population as a whole, especially in the later years of adult life. Yet, strangely enough, except in its severest forms, it rarely causes severe symptoms.

The inflammation of gastritis can be only superficial and therefore not very harmful, or it can penetrate deeply into the gastric mucosa, and in many long-standing cases cause almost complete atrophy of the gastric mucosa. In a few cases, gastritis can be very acute and severe, with ulcerative excoriation of the stomach mucosa by the stomach's own peptic secretions.

The cause of gastritis in most instances is not known. In the past this was ascribed mainly to irritant foods, but the present clinical belief is that almost no food can be as irritant to the gastric mucosa as the normal acid-pepsin gastric juices themselves. Therefore, patients with gastritis are usually told to eat almost any food that will not cause nausea or burning epigastric pain. Yet, a few substances that are often ingested can be very damaging to the protective gastric mucosal barrier—that is, to the mucous glands and the tight epithe-

hal junctions between the gastric lining cells—often leading to severe acute or chronic gastritis. The two most common of these substances are *alcohol* and *aspirin*.

The Gastric Barrier and Its Penetration in Gastritis. Absorption from the stomach is normally very slight. This low level of absorption is mainly caused by two specific features of the gastric mucosa: (1) it is lined with highly resistant mucous cells that secrete a very viscid and adherent mucus, and (2) it has very tight junctions between the adjacent epithelial cells. These two together are called the "gastric barrier." Normally this barrier is so resistant to diffusion that even the highly concentrated hydrogen ions of the gastric juice, averaging about 100,000 times the concentration of the hydrogen ions in the plasma, barely diffuse through the epithelial membrane. However, in gastritis, this barrier becomes inflamed and its permeability is greatly increased. The hydrogen ions do then diffuse into the stomach epithelium, creating additional havoc and leading to a vicious circle of progressive stomach mucosal damage and atrophy. It also makes the mucosa susceptible to peptic digestion, thus frequently resulting in gastric ulcer.

Gastric Atrophy. In many persons who have chronic gastritis, the mucosa gradually becomes atrophic until little or no gastric gland activity remains. It is also believed that some persons develop autoimmunity against the gastric mucosa, this also leading eventually to gastric atrophy. Loss of the stomach secretions in gastric atrophy leads to achlorhydria and, occasionally, to *pernicious anemia*.

Achlorhydria (and Hypochlorhydria). Achlorhydria means simply that the stomach fails to secrete hydrochloric acid, and it is diagnosed when the pH of the gastric secretions fails to decrease below 6.5 following maximal stimulation. Hypochlorhydria means diminished acid secretion. Usually, when acid is not secreted pepsin also is not secreted, and, even if it is, the lack of acid prevents it from functioning because pepsin requires an acid medium for activity. Obviously, then, essentially all digestive function in the stomach is lost when achlorhydria is present.

The method usually used to determine the degree of hypochlorhydria is to inject 0.4 milligram of pentagastrin (a synthetic gastrin—see Chapter 64) and then to aspirate the stomach secretions through a tube for the following hour. Each 15-minute sample of stomach secretion is titrated against sodium hydroxide to a pH of 7.0 using a pH meter, which gives a measure of *total acid* secreted. The maximum rate of acid secretion during any 15-minute interval in the normal person rises to almost exactly 1 milliequivalent of hydrochloric acid per minute, and the total hydrochloric acid secreted during the entire hour after the injection averages about 22 milliequivalents. All degrees of hypochlorhydria—down to no acid whatsoever—occur.

Though achlorhydria is associated with depressed or even no digestive capability by the stomach, the overall digestion of food in the entire gastrointestinal tract is still almost normal. This is because trypsin and other enzymes secreted by the pancreas are capable of digesting most of the protein in the diet—particularly if the food is well chewed so that no portion of the protein is protected by collagen fibers, which need peptic activity for most effective digestion.

Pernicious Anemia in Gastric Atrophy. Pernicious anemia is a common accompaniment of achlorhydria and gastric atrophy. The normal gastric secretions contain a glycoprotein called *intrinsic factor*, which is secreted by the same oxyntic (parietal) cells that secrete the hydrochloric acid; intrinsic factor must be present for adequate absorption of vitamin B_{12} from the ileum. The intrinsic factor combines with vitamin B_{12} and protects it from being destroyed as it passes through the gastrointestinal tract. Then, when the intrinsic factor–vitamin B_{12} complex reaches the ileum, the intrinsic factor binds with receptors on the ileal epithelial surface. In some yet unknown way this makes it possible for the vitamin B_{12} to be absorbed. In the absence of intrinsic factor, an adequate amount of vitamin B_{12} is not made available from the foods. As a result, maturation failure occurs in the bone marrow, resulting in pernicious anemia. This subject was discussed in more detail in Chapter 4.

Pernicious anemia also occurs frequently when most of the stomach has been removed for treatment of either stomach ulcer or gastric cancer or when the ileum, where vitamin B_{12} is almost entirely absorbed, is removed.

PEPTIC ULCER

A peptic ulcer is an excoriated area of the mucosa caused by the digestive action of gastric juice. Figure 66–1 illustrates the points in the gastrointestinal tract at which peptic ulcers frequently occur, showing that by far the most frequent site is in the first few centimeters of the duodenum. In addition, peptic ulcers frequently occur along the lesser curvature of the antral end of the stomach or, more rarely, in the lower end of the esophagus where stomach juices frequently reflux. A peptic ulcer called a *marginal ulcer* also frequently occurs wherever an abnormal opening, such as a gastrojejunostomy, is made between the stomach and some portion of the small intestine.

Basic Cause of Peptic Ulceration. The usual cause of peptic ulceration is an *imbalance* between the rate of secretion of gastric juice and the degree of protection afforded by the gastroduodenal mucosal barrier and the neutralization of the gastric acid by duodenal juices. It will be recalled that all areas normally exposed to gastric juice are well supplied with mucous glands, beginning with the compound mucous glands of the *lower esophagus*, then *including the mucus cell coating* of the stomach mucosa, the mucous neck cells of the gastric glands, the deep pyloric glands that secrete mainly mucus, and, finally, the glands of Brunner of the upper duodenum, which secrete a highly alkaline mucus.

In addition to the mucus protection of the mucosa, the duodenum is also protected by the alkalinity of the

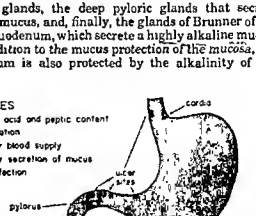


Figure 66–1. Peptic ulcer.

small intestinal secretions. Especially important is pancreatic secretion, which contains large quantities of sodium bicarbonate that neutralize the hydrochloric acid of the gastric juice, thus inactivating the pepsin to prevent digestion of the mucosa. Two additional mechanisms insure that this neutralization of gastric juices is complete:

1. When excess acid enters the duodenum, it reflexly inhibits gastric secretion and peristalsis in the stomach, both nervously and hormonally, thereby decreasing the rate of gastric emptying. This allows increased time for pancreatic secretion to enter the duodenum to neutralize the acid already present. After neutralization has taken place, the reflex subsides and more stomach contents are emptied.

2. The presence of acid in the small intestine liberates secretin from the intestinal mucosa, which then passes by way of the blood to the pancreas to promote rapid secretion of pancreatic juice containing a high concentration of sodium bicarbonate, thus making more sodium bicarbonate available for neutralization of the acid. These mechanisms were discussed in detail in Chapters 63 and 64 in relation to gastrointestinal motility and secretion.

Therefore, a peptic ulcer can be caused in either of two ways: (1) *excess secretion of acid and pepsin* by the gastric mucosa, or (2) *diminished capability of the gastroduodenal mucosal barrier to protect against the digestive properties of the acid-pepsin complex.*

Experimental Peptic Ulcer. Experimental peptic ulcers have been created in dogs and other animals in the following ways: (a) *Feeding ground glass* to an animal causes excoriation of the pyloric wall and allows the peptic juices to begin digesting the deeper layers of the mucosa. (b) *Transplantation of the pancreatic duct to the ileum* removes the normal neutralizing effect of pancreatic secretion in the duodenum and, therefore, allows the gastric juice to attack the mucosa of the upper duodenum. (c) *Repeated injection of histamine* causes excessive secretion of gastric juice. (d) *Continuol infusion of hydrochloric acid* through a tube into the stomach causes direct damage to the stomach mucosa and also prevents full neutralization of the gastric juices by the pancreatic and other secretions of the small intestine. Therefore, peptic activity is enhanced, and ulcers appear in either the stomach or duodenum. (e) A portion of the *stomach is anastomosed directly to the small intestine* so that gastric juice can pass directly and rapidly into the small intestine. The mucosa of the small intestine, except at the uppermost part of the duodenum, is not sufficiently resistant to gastric juice to prevent peptic digestion. (f) *Obstruction of the blood flow* or even reduction of the blood flow to an area of the stomach or upper duodenum will cause ulcers to develop because the local epithelium degenerates or cannot produce appropriate protective secretions.

In summary, any factor that (1) *increases the rate of production of gastric juice* or (2) *blocks the normal protective mechanisms* against this juice can produce peptic ulcers. The same general principles apply to the development of peptic ulcers in the human being.

Causes of Peptic Ulcer in the Human Being. About 85 per cent of the patients with peptic ulcer *of the duodenum* secrete approximately two times as much gastric acid following injection of a test dose of pentagastrin as do normal persons. Therefore, it is believed that duodenal ulcers are caused by excessive secretion of acid and pepsin by the gastric glands. In the 15 p cent of the patients who secrete normal amounts of ac it is presumed that one of four abnormalities is t usual cause: (a) possible secretion of an abnormal muc that has less than normal protective value, (b) dimi ished secretion of mucus, (c) failure of the normal du denal-gastric feedback mechanisms to limit the rate gastric emptying into the duodenum, or (d) failure the secretin-pancreatic and secretin–bile duct feedback mechanisms to cause the secretion of enough alkaline pancreatic juice and bile to neutralize the gastric juice as it enters the duodenum.

The development of duodenal peptic ulcers is strongly hereditary. For instance, the offspring of persons who secrete excessive amounts of gastric acid, like their parents, tend to secrete excessive amounts of acid. Also, the offspring of patients who have diminished mucosal protection have a strong hereditary tendency to develop peptic ulceration as well.

In addition to hereditary factors, psychogenic factors that cause stress anxiety seem also to play a role in peptic ulceration, though this probably occurs mainly in persons who already have a hereditary basis for developing ulcers. The number of persons who developed peptic ulcer increased greatly during the air raids of London in World War II. Therefore, it is believed that many instances of duodenal peptic ulcer in the human being are precipitated by excessive stimulation of the dorsal motor nucleus of the vagus by signals originating elsewhere in the brain, with this in turn greatly increasing the stomach secretions.

Paradoxically, *gastric* ulcers, in contradistinction to duodenal ulcers, often occur in patients who have normal or even slightly low secretion of hydrochloric acid. However, these patients frequently have an associated, often atrophic, gastritis, indicating that ulceration in the stomach most often results from reduced resistance of the stomach mucosa to digestion rather than from excess secretion of gastric juice. Stomach ulceration frequently occurs in patients who have ingested substances such as aspirin or alcohol that reduce the mucosal resistance. Also, reflux of duodenal contents into the stomach often leads to gastric ulceration; this results mainly from the bile acids in the refluxed chyme because these acids have a detergent effect that reduces the mucosal resistance.

Physiology of Treatment. The usual medical treatment for peptic ulcer is a combination of (1) reduction of stressful situations that might lead to excessive acid secretion, (2) administration of antacid drugs to neutralize much of the acid in the stomach secretions, (3) administration of the drug *cimetidine* which blocks the effect of histamine on H_2-receptors and thereby reduces the action of gastrin in stimulating gastric juice secretion by 70 to 80 per cent, (4) interdiction of smoking because statistical studies have shown that smokers are several times as prone to have peptic ulcers as are nonsmokers, and (5) removal of such ulcer-causing factors as alcohol, aspirin, or other substances that might break down the gastroduodenal mucosal barrier.

Surgical treatment of peptic ulceration is usually by one or both of two procedures: (1) *vagotomy* or (2) *removal of a portion of the stomach*. Fortunately, however, since the advent of the drug cimetidine, the more drastic surgical procedures are employed far less frequently than in the past.

Vagotomy means section of the vagus nerves to the

stomach; this temporarily blocks almost all secretion of acid and pepsin by the stomach and often cures the ulcer or ulcers within a week after the operation is performed. Unfortunately, though, a large amount of basal stomach secretion returns after a few months, and in many patients the ulcer also returns. Also, *gastric atony* usually follows section of both vagus trunks; this can be very distressing, since stomach motility is often reduced so much that gastric emptying becomes minimal, leading to partial or sometimes almost total pyloric obstruction. Several different procedures have been developed to prevent this: (1) One is *selective section of the vagus nerve branches to the oxyntic glandular regions of the stomach* while leaving intact the vagus nerve branches to the pyloric region of the stomach, to the duodenum, the pancreas, and the biliary system. In many patients this can reduce gastric acid secretion sufficiently to cure the ulceration without leading to other unwanted effects. (2) A second procedure is to perform along with the vagotomy a *pyloroplasty*, which is an operation to increase the size of the opening of the pylorus from the stomach into the duodenum Or (3) along with vagotomy, the stomach antrum and pylorus are removed, and the body of the stomach is connected directly to the proximal end of the duodenum

In the past, large numbers of patients were treated by removing the lower three fourths to four fifths of the stomach and then anastomosing the stomach to the jejunum If less than this amount of the stomach is removed, far too much gastric juice continues to be secreted, and a *marginal ulcer* soon develops where the stomach is anastomosed to the intestine. Since a marginal ulcer is usually equally as debilitating as, or sometimes even more debilitating than, the original ulcer, nothing will have been accomplished. Fortunately, this very severe procedure is rarely required in present-day therapy of peptic ulcer

DISORDERS OF THE SMALL INTESTINE

Abnormal Digestion of Food in the Small Intestine; Pancreatic Failure. Perhaps the commonest cause of abnormal digestion is failure of the pancreas to secrete its juice into the small intestine. Lack of pancreatic secretion frequently occurs (a) in *pancreatitis*, which is discussed later, (b) when the *pancreatic duct is blocked* by a gallstone at the papilla of Vater, or (c) after the *head of the pancreas has been removed* because of malignancy. Loss of pancreatic juice means loss of trypsin, chymotrypsin, carboxypolypeptidase, pancreatic amylase, pancreatic lipase, and still a few other digestive enzymes. Without these enzymes, as much as half the fat entering the small intestine may go unabsorbed and as much as one third of the proteins and starches. As a result, large portions of the ingested food are not utilized for nutrition; and copious, fatty feces are excreted.

Pancreatitis. Pancreatitis means inflammation of the pancreas, and this can occur in the form of either *acute pancreatitis* or *chronic pancreatitis*. The commonest cause of pancreatitis is alcohol and the second most common cause is blockage of the papilla of Vater by a gallstone; the two together account for over 90 per cent of all cases. When a gallstone blocks the papilla of Vater, this blocks the main secretory duct from the pancreas as well as the common bile duct. The pancreatic enzymes are then dammed up in the ducts and acini of the pancreas. Eventually, so much trypsinogen accumulates that it *overcomes the trypsin inhibitor* in the secretions, and a small quantity of trypsinogen becomes activated to form trypsin. Once this happens the trypsin activates still more trypsinogen as well as chymotrypsinogen and carboxypolypeptidase, resulting in a vicious circle until most of the proteolytic enzymes in the pancreatic ducts and acini become activated. These rapidly digest large portions of the pancreas itself, sometimes completely and permanently destroying the ability of the pancreas to secrete digestive enzymes.

Often in acute pancreatitis the proteolytic enzymes eat their way all the way to the surface of the pancreas, and pancreatic juice then empties into the peritoneal cavity where additional proteolytic activity occurs, causing *chemical peritonitis* However, as soon as large portions of the pancreas have been destroyed, the acute stage of pancreatitis subsides, leaving the person with diminished or sometimes totally absent pancreatic secretion into the gut.

Ordinarily, the islets of Langerhans are not seriously affected by pancreatitis, so that insulin still continues to be secreted by the pancreas even though secretion of pancreatic juice into the intestine is markedly reduced.

Malabsorption by the Small Intestinal Mucosa—"Sprue." Occasionally, nutrients are not adequately absorbed from the small intestine even though the food is well digested. Several different diseases can cause decreased absorbability by the mucosa; these are often classified together under the general heading of *sprue*. Obviously, also, malabsorption can occur when large portions of the small intestine have been removed.

Nontropical Sprue. One type of sprue, called variously by the names *idiopathic sprue, celiac disease* (in children), or *gluten enteropathy*, results from the toxic effects of *gluten* present in certain types of grains, especially wheat and rye It is the partly digested fractions of gluten containing glutamine that are responsible for this toxic effect, perhaps because of a direct destructive effect on the intestinal epithelial cells, or perhaps as the result of an immunological or allergic reaction In milder forms of the disease, the microvilli of the absorbing epithelial cells on the villi are destroyed, thus decreasing the absorptive surface area as much as twofold In the more severe forms, the gluten products cause early and total destruction and dissolution of newly forming epithelial cells in the crypts of Lieberkühn Therefore, these cells fail to migrate upward onto the villi. As a result, the villi themselves become blunted or disappear altogether, thus still further reducing the absorptive area of the gut. Removal of wheat and rye flour from the diet frequently results in a miraculous cure within weeks, especially in children with this disease.

Tropical Sprue. A different type of sprue called *tropical sprue* frequently occurs in the tropics and can often be treated with antibacterial agents. Even though no specific bacterium has ever been implicated as the cause, it is believed that this variety is often caused by inflammation of the intestinal mucosa resulting from yet unidentified infectious agents.

Malabsorption in Sprue. In the early stages of sprue, the absorption of fats is more impaired than the absorption of other digestive products. The fat appearing in the stools is almost entirely in the form of soaps

rather than undigested neutral fat, illustrating that the problem is one of absorption and not of digestion. In this stage of sprue, the condition is frequently called *idiopathic steatorrhea*, which means simply excess fats in the stools as a result of unknown causes

In more severe cases of sprue the absorption of proteins, carbohydrates, calcium, vitamin K, folic acid, and vitamin B_{12}, as well as many other important substances, becomes greatly impaired As a result, the person suffers (1) severe nutritional deficiency, often developing severe wasting of the body, (2) osteomalacia (demineralization of the bones because of calcium lack), (3) inadequate blood coagulation due to lack of vitamin K, and (4) macrocytic anemia of the pernicious anemia type, owing to diminished vitamin B_{12} and folic acid absorption

Regional Enteritis and Appendicitis. Regional enteritis means an inflammatory condition of the terminal portion of the ileum, and appendicitis means inflammation of the appendix Appendicitis is usually caused by an actual infection in the appendix, while the cause of regional enteritis is not known but in some instances might also result from a low grade infectious process, or perhaps an allergic or autoimmune reaction Either appendicitis or ileitis cause crampy pain that is referred to the mid-abdomen Simultaneously, each also elicits intense enterointestinal reflexes resulting in severe inhibition of gastrointestinal mobility. As a result, functional obstruction often occurs in the small bowel, causing symptoms of acute intestinal obstruction, mainly distension of the upper small intestine associated with protracted cramping pains and vomiting These will be discussed later in the chapter

DISORDERS OF THE LARGE INTESTINE

CONSTIPATION

Constipation means slow movement of feces through the large intestine, and it is often associated with large quantities of dry, hard feces in the descending colon which accumulate because of the long time available for absorption of fluid

A frequent cause of constipation is irregular bowel habits that have developed through a lifetime of inhibition of the normal defecation reflexes Newborn children are rarely constipated, but part of their training in the early years of life requires that they learn to control defecation, and this control is effected by inhibiting the natural defecation reflexes Clinical experience shows that if one fails to allow defecation to occur when the defecation reflexes are excited or if one overuses laxatives to take the place of natural bowel function, the reflexes themselves becomes progressively less strong over a period of time and the colon becomes *atonic*. For this reason, if a person establishes regular bowel habits early in life, usually defecating in the morning after breakfast when the gastrocolic and duodenocolic reflexes cause mass movements in the large intestine, the development of constipation in later life can generally be prevented

Constipation can also result from spasm of a small segment of the sigmoid. It should be recalled that motility, even normally, is weak in the large intestine, so that even a slight degree of spasm is often capable of causing serious constipation This effect frequently occurs in the so-called *irritable colon syndrome* in which bowel spasm causes constipation and consequent development of small and hard feces, a condition also often associated with crampy abdominal pains After several days, the constipation is relieved by a bowel movement that often contains a moderate to large amount of mucus. The period of constipation may then be followed by a day or so of diarrhea Following this, the cycle begins again, with repeated bouts of alternating constipation and diarrhea

Megacolon. Occasionally, a person develops constipation which is so severe that bowel movements occur only once every week or so Obviously, this allows tremendous quantities of fecal matter to accumulate in the colon, causing the colon sometimes to distend to a diameter as great as 3 to 4 inches. The condition is called megacolon, or *Hirschsprung's disease*

The most frequent cause of megarolon is lack of or deficiency of ganglion cells in the myenteric plexus in a segment of the sigmoid As a consequence, neither defecation reflexea nor peristaltic motility can occur through this area of the large intestine. And the sigmoid itself becomes small and almost spastic while feces accumulate proximal to this area, causing megacolon. Thus, the disessed portion of the large intestine appears normal by x-ray, whereas the originally normal portion of the large intestine appears greatly enlarged.

DIARRHEA

Diarrhea results from rapid movement of fecal matter through the large intestine Several causes of diarrhea with important physiological overtones are:

Enteritis. Enteritis means infection caused either by a virus or by bacteria in the intestinal tract In usual infectious diarrhea, the infection is most extensive in the large intestine and the distal end of the ileum Everywhere that the infection is present, the mucosa becomes extensively irritated, and its rate of secretion becomes greatly enhanced. In addition, the motility of the intestinal wall usually increases many fold As a result, large quantities of fluid are made available for washing the infectious agent toward the anus, and at the same time strong propulsive movements propel this fluid forward Obviously, this is an important mechanism for ridding the intestinal tract of the debilitating infection.

Of special interest is the diarrhea caused by *cholera* (and sometimes by other bacteria such as pathogenic colon bacilli) The cholera toxin directly stimulates excessive secretion of electrolytes and fluid from the crypts of Lieberkühn in the distal ileum and colon, and it specifically enhances the bicarbonate-chloride exchange mechanism, causing extreme quantities of bicarbonate ions to be secreted into the intestinal tract along with massive amounts of sodium ions and water. The loss of fluid and electrolytes can be so debilitating within a day or so that death ensues. Therefore, the most important physiological basis of therapy is simply to replace the fluid and electrolytes as rapidly as they are lost, mainly by giving the patient saline and glucose solutions. With proper therapy of this type, along with the use of antibiotics, almost no cholera patients die, but without therapy 50 per cent or more do.

Psychogenic Diarrhea. Everyone is familiar with the diarrhea that accompanies periods of nervous tension, such as during examination time or when a soldier is about to go into battle. This type of diarrhea, called psychogenic or emotional diarrhea, is caused by excessive stimulation of the parasympathetic nervous system, which greatly excites both motility and secretion of mucus in the distal colon. These two effects added together can cause marked diarrhea.

The *irritable colon syndrome*, which was discussed in relation to constipation and which often causes alternating constipation and diarrhea, is also generally believed to be at least partially a psychogenic (or emotional) disorder; it usually occurs in anxiety states.

Ulcerative Colitis. Ulcerative colitis is a disease in which extensive areas of the walls of the large intestine become inflamed and ulcerated. The motility of the ulcerated colon is often so great that mass movements occur most of the time, rather than the usual 10 to 20 minutes per day. Also, the colon's secretions are greatly enhanced. As a result, the patient has repeated diarrheal bowel movements.

The cause of ulcerative colitis is unknown. Some clinicians believe that it results from an allergic or immune destructive effect. But whatever the cause, there is a strong hereditary tendency for susceptibility to ulcerative colitis. Once the condition has progressed very far, the ulcers usually will not heal until an ileostomy is performed to allow the intestinal contents to drain to the exterior rather than to flow through the colon. Even then the ulcers sometimes fail to heal, and the only solution is removal of the colon itself.

PARALYSIS OF DEFECATION IN SPINAL CORD INJURIES

From Chapter 63 it will be recalled that defecation is normally initiated by the movement of feces into the rectum, which causes a cord-mediated defecation reflex passing from the rectum to the spinal cord and then back to the descending colon, sigmoid, rectum, and anus. This reflex adds greatly to the activity of the intrinsic defecation reflex mediated through the myenteric plexus of the sigmoid and colon wall itself. Frequently, this cord defecation reflex is blocked or altered by spinal cord injuries. For instance, destruction of the *conus medullaris* of the spinal cord destroys the sacral centers in which the cord reflex is integrated and therefore almost completely paralyzes defecation. In such instances defecation requires extensive supportive measures, such as cathartics and large enemas.

But, more frequently, the spinal cord is injured somewhere between the conus medullaris and the brain, in which case the voluntary portion of the defecation act is blocked while the basic cord reflex for defecation is still intact. Nevertheless, loss of the voluntary aid to defecation—that is, loss of the increased abdominal pressure and the stretching of the anal ring by the pelvic muscles—often makes defecation a difficult process in the person with this type of cord injury. Yet, since the cord defecation reflex can still occur, a small enema to excite the action of this reflex, usually given in the morning shortly after a meal, can often cause adequate defecation. In this way persons with spinal cord injuries can usually control their bowel movements each day.

GENERAL DISORDERS OF THE GASTROINTESTINAL TRACT

VOMITING

Vomiting is the means by which the gastrointestinal tract rids itself of its contents when almost any part of the upper gastrointestinal tract becomes excessively irritated, overdistended, or even overexcitable. Excessive distension or irritation of the duodenum provides the strongest stimulus for vomiting. Impulses are transmitted, as illustrated in Figure 66–2, by both vagal and sympathetic afferents to the bilateral *vomiting center* of the medulla, which lies near the tractus solitarius at approximately the level of the dorsal motor nucleus of the vagus. Appropriate motor reactions are then instituted to cause the vomiting act, and the motor impulses that cause the actual vomiting are transmitted from the vomiting center through the 5th, 7th, 9th, 10th, and 12th cranial nerves to the upper gastrointestinal tract and through the spinal nerves to the diaphragm and abdominal muscles.

Antiperistalsis, the Prelude to Vomiting. In the early stages of excessive gastrointestinal irritation or overdistension, *antiperistalsis* begins to occur, often many minutes before vomiting appears. The antiperistalsis may begin as far down in the intestinal tract as the ileum, and the antiperistaltic wave travels backward up the intestine at a rate of 2 to 3 cm per second; this process can actually push the intestinal contents all the

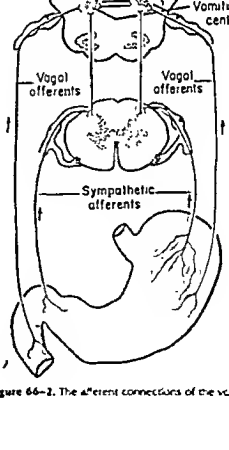


Figure 66–2. The afferent connections of the vomiting center.

way back to the duodenum and stomach within 3 to 5 minutes Then, as these upper portions of the gastrointestinal tract, especially the duodenum, become overly distended, this becomes the exciting factor that initiates the actual vomiting act. During the vomiting, strong intrinsic contractions occur in both the duodenum and the stomach, along with simultaneous relaxation of the gastroesophageal sphincter, thus causing the vomitus to be expressed into the esophagus From here a specific vomiting act involving the abdominal muscles expels the vomitus to the exterior, as explained in the following paragraph.

The Vomiting Act. Once the vomiting center has been sufficiently stimulated and the vomiting act instituted, the first effects are (1) a deep breath, (2) raising of the hyoid bone and the larynx to pull the cricoesophageal sphincter open, (3) closing of the glottis, and (4) lifting of the soft palate to close the posterior nares. Next comes a strong downward contraction of the diaphragm along with simultaneous contraction of all the abdominal muscles This obviously squeezes the stomach between the two sets of muscles, building the intragastric pressure to a high level Finally, the gastroesophageal sphincter relaxes, allowing expulsion of the gastric contents upward through the esophagus.

Thus, the vomiting act results from a squeezing action of the muscles of the abdomen associated with opening of the esophageal sphincters so that the gastric contents can be expelled

The Chemoreceptor Trigger Zone of the Medulla for Initiation of Vomiting by Drugs or by Motion Sickness. Aside from the vomiting initiated by irritative stimuli in the gastrointestinal tract itself, vomiting can also be caused by impulses arising in areas of the brain outside the vomiting center. This is particularly true of a small area located bilaterally on the floor of the fourth ventricle in or above the area postrema and called the *chemoreceptor trigger zone* Electrical stimulation of this area will initiate vomiting, but, more importantly, administration of certain drugs, including apomorphine, morphine, and some of the digitalis derivatives, can directly stimulate the chemoreceptor trigger zone and initiate vomiting Destruction of this area blocks this type of vomiting but does not block vomiting resulting from irritative stimuli in the gastrointestinal tract itself

Also, it is well known that rapidly changing motions of the body cause certain people to vomit. The mechanism for this is the following: The motion stimulates the receptors of the labyrinth, and impulses are transmitted mainly by way of the vestibular nuclei into the cerebellum, then to the chemoreceptor trigger zone, and finally to the vomiting center to cause vomiting

Cortical Excitation of Vomiting. Various psychic stimuli, including disquieting scenes, noisome odors, and other similar psychological factors, can also cause vomiting Stimulation of certain areas of the hypothalamus also causes vomiting The precise neuronal connections for these effects are not known, though it is probable that the impulses pass directly to the vomiting center and do not involve the chemoreceptor trigger zone

NAUSEA

Everyone has experienced the sensation of nausea and knows that it is often a prodrome of vomiting. Nausea is the conscious recognition of subconscious excitation in an area of the medulla closely associated with or part of the vomiting center, and it can be caused by irritative impulses coming from the gastrointestinal tract, impulses originating in the lower brain associated with motion sickness, or impulses from the cerebral cortex to initiate vomiting. However, vomiting occasionally occurs without the prodromal sensation of nausea, indicating that only certain portions of the vomiting centers are associated with the sensation of nausea

GASTROINTESTINAL OBSTRUCTION

The gastrointestinal tract can become obstructed at almost any point along its course, as illustrated in Figure 66–3; some common causes of obstruction are *cancer, fibrotic constriction resulting from ulceration or from peritoneal adhesions, spasm of a segment of the gut,* or *paralysis of a segment of the gut*

The abnormal consequences of obstruction depend on the point in the gastrointestinal tract that becomes obstructed If the obstruction occurs at the pylorus, which results often from fibrotic constriction following peptic ulceration, persistent vomiting of stomach contents occurs This obviously depresses bodily nutrition, and it also causes excessive loss of hydrogen ions from the body and can result in various degrees of alkalosis

If the obstruction is beyond the stomach, antiperistaltic reflux from the small intestine causes the intestinal juices to flow backward into the stomach, and these are vomited along with the stomach secretions In this instance the person loses large amounts of water and electrolytes so that he becomes severely dehydrated, but the loss of acids and bases may be approximately equal so that little change in acid-base balance occurs If the obstruction is near the lower end of the small intestine, then it is actually possible to vomit more basic than acidic substances; in this case acidosis may result. Also, the vomitus, after a few days of obstruction, becomes fecal in character.

Also important in obstruction of the small intestine is marked distension of the intestine proximal to the obstructed point Large quantities of fluid and electrolytes continue to be secreted into the lumen of the small intestine and even large amounts of proteins are lost *from the blood stream, partly into the intestinal lumen* and partly into the gut wall which becomes edematous as a result of the excessive distention The plasma volume diminishes because of the protein loss, and severe circulatory shock often ensues One might immediately ask. Why does the small intestine not reab-

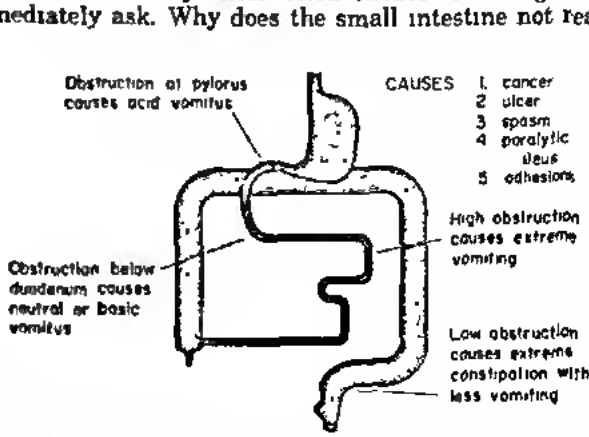


Figure 66–3. Obstruction in different parts of the gastrointestinal tract.

sorb these fluids and electrolytes? The answer is that distension of the gut usually stimulates the secretory activity of the gut but does not correspondingly increase the rate of absorption. Normally this would act to flush the chyme farther down the small intestine and therefore relieve the distension. But, if obstruction is present, obviously this normal mechanism cannot occur and instead a vicious circle of more and more distension occurs.

If the obstruction is near the distal end of the large intestine, feces can accumulate in the colon for several weeks. The patient develops an intense feeling of constipation, but in the first stages of the obstruction vomiting is not severe After the large intestine has become completely filled and it finally becomes impossible for additional chyme to move from the small intestine into the large intestine, vomiting does then begin severely. Prolonged obstruction of the large intestine will finally cause rupture of the intestine itself or dehydration and circulatory shock resulting from the severe vomiting.

CAUSES IN THE GASTROINTESTINAL TRACT AND FLATUS

Gases can enter the gastrointestinal tract from three different sources: (1) swallowed air, (2) gases formed as a result of bacterial action, and (3) gases that diffuse from the blood into the gastrointestinal tract.

Most gases in the stomach are nitrogen and oxygen derived from swallowed air, and in the normal person most of these are expelled by belching.

Only small amounts of gas are usually present in the small intestine, and much of these are air that passes from the stomach into the intestinal tract. In addition, considerable carbon dioxide also often occurs because the reaction between acidic gastric juice and bicarbonate in pancreatic juice sometimes is too rapid for the liberated CO_2 to be absorbed.

In the large intestine, the greater proportion of the gases is derived from bacterial action, including especially *carbon dioxide, methane,* and *hydrogen* These occur along with varying amounts of oxygen and nitrogen from swallowed air. When the methane and hydrogen become suitably mixed with oxygen from the swallowed air, an actual explosive mixture is occasionally formed; and use of the electric cautery during sigmoidoscopy has rarely caused colon explosions

Certain foods are known to cause greater expulsion of flatus from the large intestine than others—beans, cabbage, onion, cauliflower, corn, and certain irritant foods such as vinegar. Some of these foods serve as a suitable medium for gas-forming bacteria, especially because of unabsorbed fermentable types of carbohydrates (for instance, beans contain an undigestible sugar that passes into the colon and becomes a superior food for the colonic bacteria), but in other instances excess gas results from irritation of the large intestine, which promotes rapid expulsion of the gases before they can be absorbed.

The amount of gases entering or forming in the large intestine each day averages 7 to 10 liters, whereas the average amount expelled is usually only about 0 6 liter. The remainder is absorbed through the intestinal mucosa. Most often, a person expels large quantities of gases not because of excessive bacterial activity, but because of excessive motility of the large intestine, the gases being moved on through the large intestine before they can be absorbed.

REFERENCES

Ackerman, S H The developmental physiology of the stomach Possible contributions to the regulatory disturbances of duodenal ulcer disease *Proc Soc Exp Biol Med , 175* 119, 1984

Awouters, F., *et al* Pharmacology of antidiarrheal drugs. *Annu Rev Pharmacol Toxicol , 23* 279, 1983

Baillie, J , and Soltis, R D Systemic complications of inflammatory bowel disease *Geriatrics, 40* 53, 59, 1985

Berglunch, T The mammalian gastric parietal cell in vitro *Annu Rev Physiol , 46* 377, 1984

Blair, E (ed) Markowitz's Experimental Surgery, 6th Ed Baltimore, Williams & Wilkins, 1975

Brooks, B N Crohn's Disease Aetiology, Clinical Manifestations and Management New York, Oxford University Press, 1977

Brooks, F P (ed) Gastrointestinal Pathophysiology New York, Oxford University Press, 1974

Castell, D O Calcium channel blocking agents for gastrointestinal disorders *Am J Cardiol , 55* 210B, 1985

Cook, G C Tropical Gastroenterology New York, Oxford University Press, 1980

Duthie, H L (ed) Gastrointestinal Motility in Health and Disease Baltimore University Park Press, 1978

Eisenberg, M M Physiologic Approach to the Surgical Managament of Duodenal Ulcer Chicago, Year Book Medical Publishers, 1977

Eisenberg, M M Ulcers Boston, G K Hall, 1979

Giannella, R A Pathogenesis of acute bacterial diarrheal disorders *Annu Rev Med , 32* 341, 1981

Goodman, M J , and Sparberg, M Ulcerative Colitis New York, John Wiley & Sons, 1978

Guth, P H Pathogenesis of gastric mucosal injury *Annu Rev Med , 32* 183, 1982

Halsted, C H Intestinal absorption and malabsorption of folates. *Annu Rev Med , 31* 79, 1980

Jones, F A , *et al* Peptic Ulcer Healing Recent Studies on Carbenoxolone Baltimore, University Park Press, 1978

Kestenbaum, B , and Behar, J Pathogenesis, diagnosis, and management of reflux esophagitis *Annu Rev Med , 32* 443, 1981

Kirsner, J B , and Shorter, R O (eds) Inflammatory Bowel Disease, 2nd Ed Philadelphia, Lea & Febiger, 1980

Lebenthal, E (ed) Digestive Disease in Children New York, Orune & Stratton, 1978

Maingot, R Abdominal Operations New York, Appleton Century-Crofts, 1979

McCarthy, D M Zollinger Ellison syndrome *Annu Rev Med 33* 197, 1982

McNicholl, B , *et al* (eds.) Perspectives in Coeliac Disease Baltimore, University Park Press, 1978

Najarian, J S , and Delaney, J P (eds.) Gastrointestinal Surgery Chicago, Year Book Medical Publishers, 1979

Powell, L W , and Piper, D W (eds) Fundamentals of Gastroenterology, 3rd Ed Baltimore, University Park Press, 1979

Rankow, R M , and Polayes, I M Diseases of the Salivary Glands Philadelphia, W. B Saunders Co , 1976

Shearman, D J (ed) Diseases of the Gastrointestinal Tract and Liver New York, Churchill Livingstone, 1982

Silen, W (rev) Cope's Early Diagnosis of the Acute Abdomen, New York, Oxford University Press, 1979

Silva, J , Jr , and Fekety, R· Clostridia and antimicrobial enterocolitis. *Annu Rev Med , 32* 327, 1981.

Siesenger, M H , and Fordtran, J S.. Gastrointestinal Disease Pathophysiology, Diagnosis, Management. Philadelphia, W B Saunders Co , 1977

Strickland, R. G., and Jewell, D P · Immunoregulatory mechanisms in nonspecific inflammatory bowel disease *Annu Rev. Med., 34* 195, 1983

Welch, C E , *et al*: Manual of Lower Gastrointestinal Surgery. New York, Springer-Verlag, 1979.

XII

METABOLISM AND TEMPERATURE REGULATION

67 ■ Metabolism of Carbohydrates and Formation of Adenosine Triphosphate

68 ■ Lipid Metabolism

69 ■ Protein Metabolism

70 ■ The Liver as an Organ

71 ■ Energetics and Metabolic Rate

72 ■ Body Temperature, Temperature Regulation, and Fever

67

Metabolism of Carbohydrates and Formation of Adenosine Triphosphate

The next few chapters deal with metabolism in the body, which means the chemical processes that make it possible for the cells to continue living. It is not the purpose of this textbook, however, to present the chemical details of all the various cellular reactions, for this lies in the discipline of biochemistry. Instead, these chapters are devoted to (1) a review of the principal chemical processes of the cell, and (2) an analysis of their physiological implications, especially in relation to the manner in which they fit into the overall concept of homeostasis.

RELEASE OF ENERGY FROM FOODS AND THE CONCEPT OF "FREE ENERGY"

A great proportion of the chemical reactions in the cells is concerned with making the energy in foods available to the various physiological systems of the cell. For instance, energy is required for (a) muscular activity, (b) secretion by the glands, (c) maintenance of membrane potentials by the nerve and muscle fibers, (d) synthesis of substances in the cells, (e) absorption of foods from the gastrointestinal tract, and (f) many other functions.

Coupled Reactions. All the energy foods—carbohydrates, fats and proteins—can be oxidized in the cells, and in this process large amounts of energy are released. These same foods can also be burned with pure oxygen outside the body in an actual fire, again releasing large amounts of energy. However, this time the energy is released suddenly, all in the form of heat. The energy needed by the physiological processes of the cells is not heat but instead energy (a) to cause mechanical movement in the case of muscle function, (b) to concentrate solutes in the case of glandular secretion, and (c) to effect other functions. To provide this energy, the chemical reactions must be "coupled" with the systems responsible for these physiological functions. This coupling is accomplished by special cellular enzyme and energy transfer systems, some of which will be explained in this and subsequent chapters.

"Free Energy." The amount of energy liberated by complete oxidation of a food is called *the free energy of the food*, and this is generally represented by the symbol ΔF. Free energy is usually expressed in terms of calories per mole of food substance. For instance, the amount of free energy liberated by oxidation of 1 mole of glucose (180 grams of glucose) is 686,000 calories.

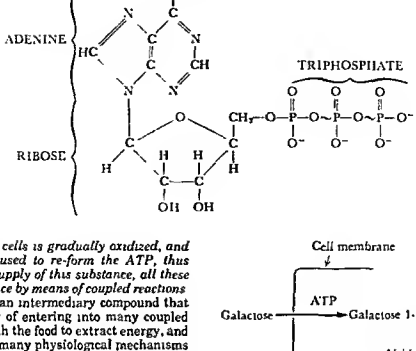
ROLE OF ADENOSINE TRIPHOSPHATE (ATP) IN METABOLISM

ATP is a labile chemical compound that is present in all cells and has the chemical structure shown in the formula on the following page.

From this formula it can be seen that ATP is a combination of adenine, ribose, and three phosphate radicals. The last two phosphate radicals are connected with the remainder of the molecule by so-called *high energy bonds*, which are indicated by the symbol ~. The amount of free energy in each of these high energy bonds per mole of ATP is approximately 7300 calories under standard conditions and about 12,000 calories under the conditions of temperature and concentrations of the reactants in the body. Therefore, removal of each phosphate radical liberates 12,000 calories of energy. After loss of one phosphate radical from ATP, the compound becomes *adenosine diphosphate* (ADP), and after loss of the second phosphate radical the compound becomes *adenosine monophosphate* (AMP). The interconversions between ATP, ADP, and AMP are the following:

$$\text{ATP} \underset{+12{,}000\text{ cal}}{\overset{-12{,}000\text{ cal}}{\rightleftharpoons}} \left\{\begin{matrix}\text{ADP}\\+\\\text{PO}_4\end{matrix}\right\} \underset{+12{,}000\text{ cal}}{\overset{-12{,}000\text{ cal}}{\rightleftharpoons}} \left\{\begin{matrix}\text{AMP}\\+\\2\text{PO}_4\end{matrix}\right\}$$

ATP is present everywhere in the cytoplasm and nucleoplasm of all cells, and essentially all the physiological mechanisms that require energy for operation obtain this directly from the ATP (or some other similar high energy compound—guanosine triphosphate, GTP, for example).



In turn, the food in the cells is gradually oxidized, and the released energy is used to re-form the ATP, thus always maintaining a supply of this substance, all these energy transfers take place by means of coupled reactions

In summary, ATP is an intermediary compound that has the peculiar ability of entering into many coupled reactions—reactions with the food to extract energy, and reactions in relation to many physiological mechanisms to provide energy for their operation For this reason, ATP has frequently been called the energy *currency* of the body that can be gained and spent again and again

The principal purpose of the present chapter is to explain how the energy from carbohydrates can be used to form ATP in the cells. At least 99 per cent of all the carbohydrates utilized by the body is used for this purpose.

CENTRAL ROLE OF GLUCOSE IN CARBOHYDRATE METABOLISM

From Chapter 65 it will be recalled that the final products of carbohydrate digestion in the alimentary tract are almost entirely glucose, fructose, and galactose—with glucose representing, on the average, about 80 per cent of these. After absorption from the intestinal tract, much of the fructose and almost all the galactose are also rapidly converted into glucose Therefore, very little fructose and galactose are present in the circulating blood. Glucose thus becomes the final common pathway for transport of almost all carbohydrates to the tissue cells.

In the case of fructose, some of this is converted into glucose as it is absorbed through the intestinal epithelial cells into the portal blood. This conversion was discussed in Chapter 65. Much of the remaining fructose and essentially all the galactose are then converted into glucose by the liver as will be explained.

In liver cells, appropriate enzymes are available to promote interconversions among the monosaccharides, as shown in Figure 67–1 Furthermore, the dynamics of the reactions are such that when the liver releases the monosaccharides back into the blood the final product is almost entirely glucose The reason for this is that the liver cells contain large amounts of *glucose phosphatase;* therefore, glucose 6-phosphate can be degraded back to glucose and phosphate, and the glucose transported back into the blood.

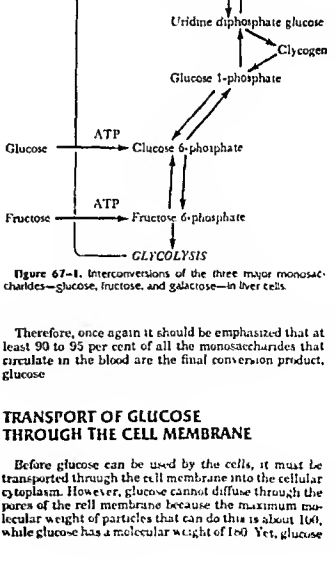


Figure 67–1. Interconversions of the three major monosaccharides—glucose, fructose, and galactose—in liver cells.

Therefore, once again it should be emphasized that at least 90 to 95 per cent of all the monosaccharides that circulate in the blood are the final conversion product, glucose

TRANSPORT OF GLUCOSE THROUGH THE CELL MEMBRANE

Before glucose can be used by the cells, it must be transported through the cell membrane into the cellular cytoplasm. However, glucose cannot diffuse through the pores of the cell membrane because the maximum molecular weight of particles that can do this is about 100, while glucose has a molecular weight of 180 Yet, glucose

does pass to the interior of the cells with a reasonable degree of freedom but is transported through the membrane by the mechanism of *facilitated diffusion*. The principles of this type of transport were discussed in Chapter 9. Basically, they are the following: In the lipid matrix of the cell membrane are large numbers of protein carrier molecules that can bind with glucose. In this bound form the glucose can be transported by the carrier from one side of the membrane to the other side and then released. Therefore, if the concentration of glucose is greater on one side of the membrane than on the other side, more glucose will be transported from the high concentration area than in the opposite direction.

It should be noted that this transport of glucose through the membranes of most tissue cells is quite different from that which occurs through the gastrointestinal membrane or through the epithelium of the renal tubules. In both these cases, the glucose is transported by the mechanism of active sodium cotransport, in which active transport of sodium provides energy for absorbing glucose against a concentration difference. This sodium cotransport mechanism functions only in certain special epithelial cells that are specifically adapted for active absorption of glucose. Instead, at all other cell membranes, glucose is transported only from higher concentration toward lower concentration by the process of facilitated diffusion, made possible by the special binding properties of the glucose carrier.

FACILITATION OF GLUCOSE TRANSPORT BY INSULIN

The rate of glucose transport, and also transport of some other monosaccharides, is greatly increased by insulin. When large amounts of insulin are secreted by the pancreas, the rate of glucose transport into most cells increases to 10 or more times the rate of transport when no insulin at all is secreted. The amounts of glucose that can diffuse to the insides of most cells of the body in the absence of insulin, with the exceptions of the liver and brain cells, are far too little to supply anywhere near the amount of glucose normally required for energy metabolism. Therefore, in effect, the rate of carbohydrate utilization by most cells is controlled by the rate of insulin secretion in the pancreas. The functions of insulin and its control of carbohydrate metabolism will be discussed in detail in Chapter 78.

Failure of Disaccharides to Be Transported. Very minute amounts of disaccharides are absorbed into the blood from the gastrointestinal tract, but essentially none of these can be transported into the cells. Therefore, no disaccharides or larger polysaccharides are utilized for cellular metabolism. Instead, they are excreted completely in the urine.

PHOSPHORYLATION OF GLUCOSE

Immediately upon entry into the cells, glucose combines with a phosphate radical in accordance with the following reaction.

$$\text{Glucose} \xrightarrow[+\,\text{ATP}]{\text{glucokinase or hexokinase}} \text{Glucose 6-phosphate}$$

This phosphorylation is promoted by the enzyme *glucokinase* in the liver or hexokinase in most other cells.

The phosphorylation of glucose is almost completely irreversible except in the liver cells, the renal tubular epithelium, and the intestinal epithelial cells in which glucose phosphatase is available for reversing the reaction. Therefore, in most tissues of the body phosphorylation serves to *capture* the glucose in the cell—once in the cell the glucose will not diffuse back out except from those special cells listed earlier that have the necessary phosphatase.

STORAGE OF GLYCOGEN IN LIVER AND MUSCLE

After absorption into the cells, glucose can be used immediately for release of energy to the cells or it can be stored in the form of *glycogen*, which is a large polymer of glucose.

All cells of the body are capable of storing at least some glycogen, but certain cells can store large amounts, especially the liver cells, which can store up to 5 to 8 per cent of their weight as glycogen, and muscle cells, which can store up to 1 to 3 per cent glycogen. The glycogen molecules can be polymerized to almost any molecular weight, the average molecular weight being 5,000,000 or greater, most of the glycogen precipitates in the form of solid granules. This conversion of the monosaccharides into a high molecular weight, precipitated compound makes it possible to store large quantities of carbohydrates without significantly altering the osmotic pressure of the intracellular fluids. Obviously, high concentrations of low molecular weight, soluble monosaccharides would play havoc with the osmotic relationships between intracellular and extracellular fluids.

GLYCOGENESIS

Glycogenesis is the process of glycogen formation, the chemical reactions for which are illustrated in Figure 67–2. From this figure it can be seen that *glucose 6-phosphate* first becomes *glucose 1-phosphate;* then this is converted to *uridine diphosphate glucose*, which is then converted into glycogen. Several specific enzymes

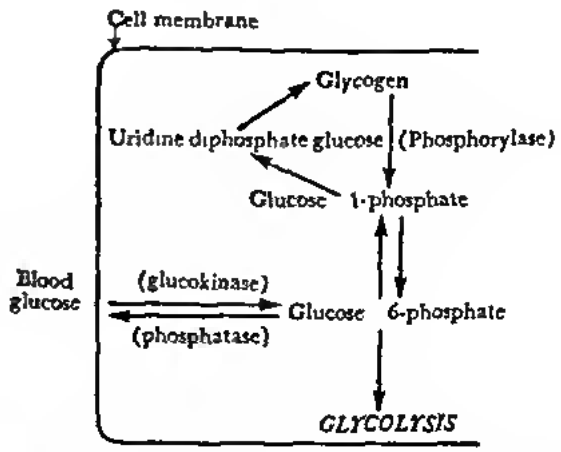


Figure 67–2. The chemical reactions of glycogenesis and glycogenolysis, showing also the interconversions between blood glucose and liver glycogen. (The phosphatase required for release of glucose from the cell is absent in muscle cells.)

are required to cause these conversions, and any monosaccharide that can be converted into glucose obviously can enter into the reactions. Certain smaller compounds, including *lactic acid, glycerol, pyruvic acid,* and *some deaminated amino acids,* can also be converted into glucose or closely allied compounds and thence into glycogen.

GLYCOGENOLYSIS

Glycogenolysis means the breakdown of glycogen to re-form glucose in the cells. Glycogenolysis does not occur by reversal of the same chemical reactions that serve to form glycogen; instead, each succeeding glucose molecule on each branch of the glycogen polymer is split away by a process of *phosphorylation,* catalyzed by the enzyme *phosphorylase* (several other enzymes split the glycogen molecule at the branching points).

Under resting conditions, the phosphorylase is in an inactive form so that glycogen can be stored but not reconverted into glucose. When it is necessary to re-form glucose from glycogen, therefore, the phosphorylase must first be activated. This is accomplished in the following two ways:

Activation of Phosphorylase by Epinephrine and Glucagon. Two hormones, epinephrine and glucagon, can specifically activate phosphorylase and thereby cause rapid glycogenolysis The initial effect of each of these hormones is to increase the formation of *cyclic adenosine monophosphate (cAMP)* in the cells. This substance then initiates a cascade of chemical reactions that activates the phosphorylase. This will be discussed in more detail in Chapter 78

Epinephrine is released by the adrenal medullae when the sympathetic nervous system is stimulated. Therefore, one of the functions of the sympathetic nervous system is to increase the availability of glucose for rapid metabolism. This function of epinephrine occurs markedly both in liver cells and in muscle, thereby contributing, along with other effects of sympathetic stimulation, to preparing the body for action, as was discussed in Chapter 57.

Glucagon is a hormone secreted by the *alpha cells* of the pancreas when the blood glucose concentration falls too low. It stimulates the formation of cAMP mainly in the liver. Therefore, its effect is primarily to convert liver glycogen into glucose and to release this into the blood, thereby elevating blood glucose concentration. The function of glucagon in blood glucose regulation is discussed in Chapter 78.

RELEASE OF ENERGY FROM THE GLUCOSE MOLECULE BY THE GLYCOLYTIC PATHWAY

Since complete oxidation of 1 gram-mole of glucose releases 686,000 calories of energy, and only 12,000 calories of energy are required to form 1 gram-mole of adenosine triphosphate (ATP), it would be extremely wasteful of energy if glucose should be decomposed all the way into water and carbon dioxide at once while forming only a single ATP molecule. Fortunately, cells contain an extensive series of different protein enzymes that cause the glucose molecule to split a little at a time in many successive steps, with its energy released in small packets to form one molecule of ATP at a time, forming a total of 38 moles of ATP for each mole of glucose utilized by the cells.

The purpose of the present section is to describe the basic principles by which the glucose molecule is progressively dissected and its energy released to form ATP.

GLYCOLYSIS AND THE FORMATION OF PYRUVIC ACID

By far the most important means by which energy is released from the glucose molecule is by the process of *glycolysis* and then *oxidation of the end-products of glycolysis.* Glycolysis means splitting of the glucose molecule to form *two molecules of pyruvic acid* This occurs by 10 successive steps of chemical reactions, as illustrated in Figure 67–3 Each step is catalyzed by at

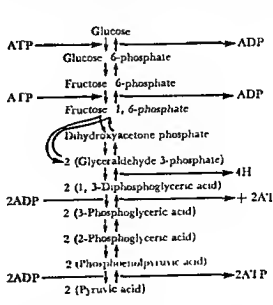


Figure 67–3. The sequence of chemical reactions responsible for glycolysis.

Net reaction:

$$\text{Glucose} + 2\text{ADP} + 2\text{PO}_4^{---} \longrightarrow 2\ \text{Pyruvic acid} + 2\text{ATP} + 4\text{H}$$

least one specific protein enzyme. Note that glucose is first converted into fructose 1,6-phosphate and then split into two three-carbon atom molecules, each of which is then converted through five successive steps into pyruvic acid.

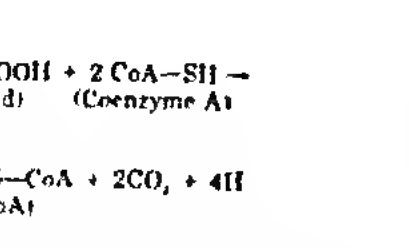
Formation of Adenosine Triphosphate (ATP) During Glycolysis. Despite the many chemical reactions in the glycolytic series, only a small portion of the free energy in the glucose molecule is released. However, between the 1,3-diphosphoglyceric acid and the 3-phosphoglyceric acid stages, and again between the phosphopyruvic acid and the pyruvic acid stages, the packets of energy released are greater than 12,000 calories per mole, the amount required to form ATP, and the reactions are coupled in such a way that ATP is formed. Thus, a total of 4 moles of ATP are formed for each mole of fructose 1,6 phosphate that is split into pyruvic acid.

Yet 2 moles of ATP had been required to phosphorylate the original glucose to form fructose 1,6-phosphate before glycolysis could begin. Therefore, the net gain in ATP molecules by the entire glycolytic process is only 2 moles for each mole of glucose utilized. This amounts to 24,000 calories of energy stored in the form of ATP, but during glycolysis a total of 56,000 calories of energy is lost from the original glucose, giving an overall *efficiency* for ATP formation of 43 per cent. The remaining 57 per cent of the energy is lost in the form of heat.

Release of Hydrogen Atoms During Glycolysis. Note in Figure 67–3 that 2 hydrogen atoms are released from each molecule of glyceraldehyde during its conversion to 1,3-diphosphoglyceric acid. And, since 2 molecules of glyceraldehyde are formed from each glucose molecule, a total of 4 hydrogen atoms is released. Later in the chapter we will see that still many more hydrogen atoms are released when pyruvic acid is split into its component parts and that these hydrogen atoms plus the 4 released during glycolysis are oxidized to provide most of the energy that is eventually used in the synthesis of ATP.

CONVERSION OF PYRUVIC ACID TO ACETYLCOENZYME A

The next stage in the degradation of glucose is (1) facilitated transport of the 2 derivative pyruvic acid molecules into the matrix of the mitochondria, and then (2) conversion of these into 2 molecules of *acetylcoenzyme A* (acetyl CoA) in accordance with the following reaction:



From this reaction it can be seen that 2 carbon dioxide molecules and 4 hydrogen atoms are released, while the remainders of the 2 pyruvic acid molecules combine with coenzyme A, a derivative of the vitamin pantothenic acid, to form 2 molecules of acetyl-CoA. In this conversion, no ATP is formed but up to 6 molecules of ATP are formed when the 4 hydrogen atoms are later oxidized, as will be discussed in a later section.

THE CITRIC ACID CYCLE

The next stage in the degradation of the glucose molecule is called the *citric acid cycle* (also called the *tricarboxylic acid cycle* or *Krebs cycle*). This is a sequence of chemical reactions in which the acetyl portion of acetyl-CoA is degraded to carbon dioxide and hydrogen atoms. These reactions all occur in the matrix of the mitochondrion. The released hydrogen atoms are subsequently oxidized, as will be discussed later, releasing tremendous amounts of energy to form ATP.

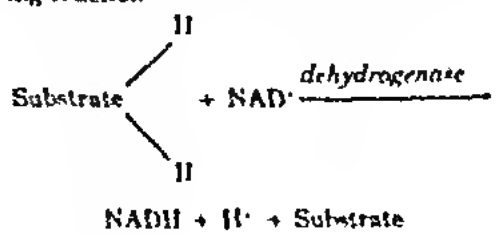
Figure 67–4 shows the different stages of the chemical reactions in the citric acid cycle. The substances to the left are added during the chemical reactions, and the products of the chemical reactions are shown to the right. Note at the top of the column that the cycle begins with *oxaloacetic acid*, and then at the bottom of the chain of reactions, *oxaloacetic acid* is formed once again. Thus, the cycle can continue over and over.

In the initial stage of the citric acid cycle, *acetyl CoA combines with oxaloacetic acid to form citric acid*. The coenzyme A portion of the acetyl-CoA is released and can be used again and again for the formation of still more quantities of acetyl-CoA from pyruvic acid. The acetyl portion, however, becomes an integral part of the citric acid molecule. During the successive stages of the citric acid cycle, several molecules of water are added, and *carbon dioxide* and *hydrogen atoms* are released at various stages in the cycle, as shown on the right in the figure.

The net results of the entire citric acid cycle are given in the legend of Figure 67–4, illustrating that for each molecule of glucose originally metabolized, 2 acetyl-CoA molecules enter into the citric acid cycle along with 6 molecules of water. These then are degraded into 4 carbon dioxide molecules, 16 hydrogen atoms, and 2 molecules of coenzyme A.

Formation of ATP in the Citric Acid Cycle. Not a great amount of energy is released during the citric acid cycle itself, in only one of the chemical reactions—during the change from α ketoglutaric acid to succinic acid—is a molecule of ATP formed. Thus, for each molecule of glucose metabolized, 2 acetyl-CoA molecules pass through the citric acid cycle, each forming a molecule of ATP, or, a total of 2 molecules of ATP is formed.

Function of the Dehydrogenases and Nicotinamide Adenine Dinucleotide (NAD^+) in Causing Release of Hydrogen Atoms. As noted at several points in this discussion, hydrogen atoms are released during different chemical reactions—4 hydrogen atoms during glycolysis, 4 during the formation of acetyl CoA from pyruvic acid, and 16 in the citric acid cycle, this makes a total of 24 hydrogen atoms. However, the hydrogen atoms are not simply turned loose in the intracellular fluid. Instead, they are released in packets of two, and in each instance the release is catalyzed by a specific protein enzyme called a *dehydrogenase*. Twenty of the 24 hydrogen atoms immediately combine with nicotinamide adenine dinucleotide (NAD^+), a derivative of the vitamin niacin, in accordance with the following reaction:



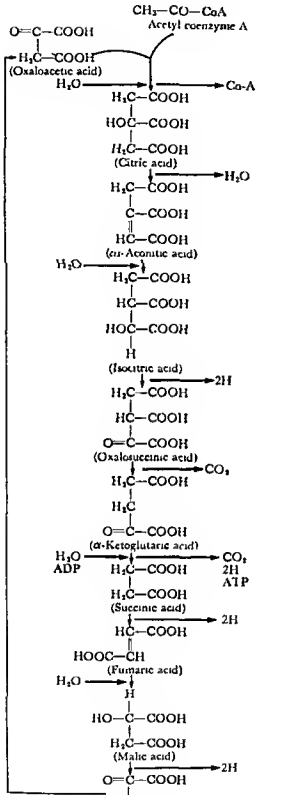


Figure 67–4. The chemical reactions of the citric acid cycle, showing the release of carbon dioxide and an especially large number of hydrogen atoms during the cycle

Net reaction per molecule of glucose:

$$2\ \text{Acetyl-CoA} + 6H_2O + 2ADP \longrightarrow 4CO_2 + 16H + 2CoA + 2ATP$$

This reaction will not occur without the initial intermediation of the specific dehydrogenase nor without the availability of NAD^+ to act as a hydrogen carrier. Both the free hydrogen ion and the hydrogen bound with NAD^+ subsequently enter into the oxidative chemical reactions that form tremendous quantities of ATP, as will be discussed later.

The remaining 4 hydrogen atoms released during the breakdown of glucose—the four released during the citric acid cycle between the succinic and fumaric acid stages—combine with a specific dehydrogenase but are not then subsequently released to NAD^+. Instead, they pass directly from the dehydrogenase into the oxidative processes.

Function of Decarboxylases in Causing Release of Carbon Dioxide. Referring again to the chemical reactions of the citric acid cycle and also to those for formation of acetyl-CoA from pyruvic acid, we find that there are three stages in which carbon dioxide is released. To cause the carbon dioxide release, other specific protein enzymes, called *decarboxylases*, split the carbon dioxide away from the substrate. The carbon dioxide in turn becomes dissolved in the fluids of the cells and is transported to the lungs where it is expired from the body (see Chapter 41).

FORMATION OF ATP BY OXIDATIVE PHOSPHORYLATION

Despite all the complexities of (a) glycolysis, (b) the citric acid cycle, (c) dehydrogenation, and (d) decarboxylation, pitifully small amounts of ATP are formed during all these processes—only 2 ATP molecules in the glycolysis scheme and another 2 in the citric acid cycle. Instead, 90 per cent of the total ATP formed by glucose metabolism is formed during subsequent oxidation of the hydrogen atoms that are released during these earlier stages of glucose degradation. Indeed, the principal function of all these earlier stages is to make the hydrogen of the glucose molecule available in forms that can be utilized for oxidation.

Oxidation of hydrogen is accomplished by a series of enzymatically catalyzed reactions in the mitochondria that (a) split each hydrogen atom into a hydrogen ion and an electron, and (b) use the electrons eventually to change the dissolved oxygen of the fluids into hydroxyl ions. Then the hydrogen and hydroxyl ions combine with each other to form water. During the sequence of oxidative reactions, tremendous quantities of energy are released to form ATP. Formation of ATP in this manner is called *oxidative phosphorylation*. This occurs entirely in the mitochondria by a highly specialized process called the chemiosmotic mechanism.

The Chemiosmotic Mechanism of the Mitochondria

Ionization of Hydrogen, the Electron Transport Chain, and Formation of Water. The first step in oxidative phosphorylation in the mitochondria is to ionize the hydrogen atoms that have been removed from the food substrates. As described earlier, these hydrogen atoms are removed in pairs: one immediately becomes a hydrogen ion, H^+, and the other combines with NAD^+

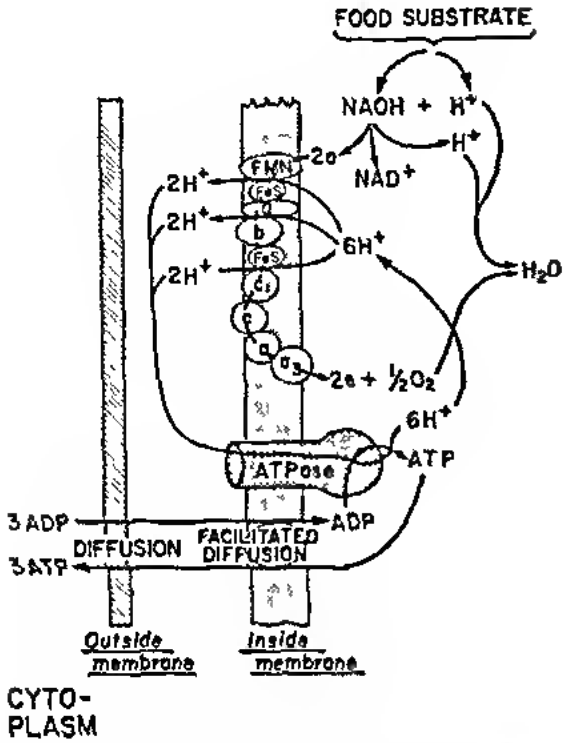


Figure 67–5. The chemiosmotic theory of oxidative phosphorylation for forming great quantities of ATP

to form NADH The upper portion of Figure 67–5 shows in color the subsequent fate of the NADH and H^+ The initial effect is to release the other hydrogen atom bound with NAD to form another hydrogen ion, H^+; this process also reconstitutes NAD^+ that will be reused again and again

The electrons that are removed from the hydrogen atoms to cause the ionization immediately enter an *electron transport chain* that is an integral part of the inner membrane (the shelf membrane) of the mitochondrion This transport chain consists of a series of electron acceptors that can be reversibly reduced or oxidized by accepting or giving up electrons The important members of this electron transport chain include *flavoprotein, several iron sulfide proteins, ubiquinone,* and *cytochromes B, C_1, C, A, and A_3* Each electron is shuttled from one of these acceptors to the next until it finally reaches cytochrome A_3, which is called *cytochrome oxidase* because it is capable, by giving up two electrons, of causing elemental oxygen to combine with hydrogen ions to form water

Thus, Figure 67–5 illustrates transport of electcons through the electron chain and then ultimate use of these by cytochrome oxidase to cause the formation of water molecules During the transport of these electrons through the electron transport chain energy is released that is later used to cause synthesis of ATP.

Pumping of Hydrogen Ions into the Outer Chamber of the Mitochondrion, Caused by the Electron Transport Chain. As the electrons pass through the electron transport chain, large amounts of energy are released. This energy is used to pump hydcogen ions from the inner matrix of the mitochondrion into the outer chamber between the inner and outer mitochondrial membranes This creates a high concentration of hydrogen ions in this chamber, and it also creates a [illegible]ive electrical potential in the inner matrix.

Formation of ATP. The next step in oxidative phosphorylation is to convert ADP into ATP. This occurs in conjunction with a large protein molecule that protrudes all the way through the inner mitochondrial membrane and projects with a knoblike head into the inner matrix. This molecule is an ATPase, the physical nature of which is illustrated in Figure 67–5. It is called *ATP synthetase* It is believed that the high concentration of hydrogen ions in the outer chamber and the large electrical potential difference across the inner membrane cause the hydrogen ions to flow into the mitochondrial matrix *through the substance of the ATPase molecule* In doing so, energy derived from this hydrogen ion flow is utilized by the ATPase to convert ADP into ATP by combining ADP with inorganic phosphate (Pi), thus forming an additional high energy phosphate bond

The final step in the process is transfer of the ATP fcom the inside of the mitochondrion back to the cytoplasm This occurs by facilitated diffusion outward through the inner membrane and then by simple diffusion through the very permeable outer mitochondrial membrane In turn, ADP is rontinually transferred in a like manner back into the mitochondrial matrix for continual conversion into ATP.

For each 2 electrons that pass through the entire electron transport chain (representing the ionization of 2 hydrogen atoms), up to 3 ATP molecules are synthesized

SUMMARY OF ATP FORMATION DURING THE BREAKDOWN OF GLUCOSE

We can now determine the total number of ATP molecules formed by the energy from 1 molecule of glurose

First, during glyrolysis, 4 molecules of ATP are formed while 2 are expended to cause the initial phosphorylation of glucose to start the process going. This gives a net gain of *2 molecules of ATP.*

Second, during each revolution of the citric acid cycle, 1 molecule of ATP is formed. However, because each glucose molecule splits into 2 pyruvic acid molecules, there are two revolutions of the cycle for each molecule of glucose metabolized, giving a net production of *2 molecules of ATP*

Third, during the entire schema of glucose breakdown, a total of 24 hydrogen atoms is released during glycolysis and during the citric acid cycle. Twenty of these atoms are oxidized by the schema of oxidative reactions shown in Figure 67–5, with the release of up to 3 ATP molecules per 2 atoms of hydrogen metabolized This gives a production of *30 ATP molecules.*

Fourth, the remaining 4 hydrogen atoms are released by their dehydcogenase into the oxidative schema in the mitochondrion beyond the first stage of Figure 67–5, so that for these 4 hydrogen atoms only 2 ATP molecules are usually released for each 2 hydrogen atoms oxidized, giving a total of *4 ATP molecules.*

Now, adding all the ATP molecules formed, we find a maximum of *38 ATP molecules* formed for each molecule of glucose degraded to carbon dioxide and water. Thus, 456,000 calories of energy can be stored in the form of ATP, while 686,000 calories are released during the complete oxidation of each gram-mole of glucose. This represents an overall maximum *efficiency* of energy transfer of 66 per cent. The remaining 34 per cent of the energy becomes heat and therefore cannot be used by the cells to perform specific functions

CONTROL OF GLYCOLYSIS AND OXIDATION BY ATP AND ADP

Continual release of energy from glucose when the energy is not needed by the cells would be an extremely wasteful process. Fortunately, glycolysis and the subsequent oxidation of hydrogen atoms are continually controlled in accordance with the needs of the cells for ATP. This control is accomplished by a multitude of feedback control mechanisms within the chemical schemas. However, among the more important of these are the effects of both ADP and ATP in controlling the rates of chemical reactions in the energy metabolism sequence.

An important way in which ATP helps control energy metabolism is its effect in causing an *allosteric inhibition of the enzyme phosphofructokinase* Because this enzyme promotes the formation of fructose 1,6-phosphate, one of the initial steps in the glycolytic series of reactions, the net effect of excess ATP, therefore, is to stop glycolysis, which in turn stops most carbohydrate metabolism Conversely, ADP (and AMP as well) causes the opposite allosteric change in this enzyme, greatly increasing its activity. Therefore, whenever ATP is used by the tissues for energy, this reduces the ATP inhibition of the enzyme but at the same time increases its activity as a result of the ADP that is formed, and the glycolytic process is set in motion Soon the total cellular store of ATP is replenished.

Another control linkage is the *citrate ion* formed in the citric acid cycle An excess of this ion also *strongly inhibits phosphofructokinase*, thus preventing the glycolytic process from getting ahead of the ability of the citric acid cycle to utilize the pyruvic acid formed during glycolysis.

A third way by which the ATP-ADP-AMP system controls carbohydrate metabolism, as well as controls energy release from fats and proteins, is the following. Referring back to the various chemical reactions for energy release, we see that at many different stages ADP is converted into ATP. If ADP is not available, the reactions at each of these stages simply cannot occur, thus stopping the degradation of the glucose molecule, fat molecule, or protein molecule Then, when ATP is used to energize the different physiological functions in the cell, the resulting ADP and AMP turn on the energy processes again, and the ADP and AMP are almost instantly returned to the ATP state In this way, essentially a full store of ATP is automatically maintained all the time, except during extreme cellular activity as might occur in very heavy exercise.

ANAEROBIC RELEASE OF ENERGY—"ANAEROBIC GLYCOLYSIS"

Occasionally, oxygen becomes either unavailable or insufficient so that oxidative phosphorylation cannot take place. Yet, even under these conditions, a small amount of energy can still be released to the cells by glycolysis, for the chemical reactions in the glycolytic breakdown of glucose to pyruvic acid do not require oxygen. Unfortunately, this process is extremely wasteful of glucose because only 24,000 calories of energy are used to form ATP for each molecule of glucose utilized, which represents only a little over 3 per cent of the total energy in the glucose molecule. Nevertheless, this release of glycolytic energy to the cells can be a life-saving measure for a few minutes when oxygen becomes unavailable.

Formation of Lactic Acid During Anaerobic Glycolysis. The *law of mass action* states that as the end-products of a chemical reaction build up in the reacting medium, the rate of the reaction approaches zero. The two end-products of the glycolytic reactions (see Fig. 67–3) are (1) pyruvic acid and (2) hydrogen atoms, which are combined with NAD^+ to form NADH and H^+. The buildup of either or both of these would stop the glycolytic process and prevent further formation of ATP. Fortunately, when their quantities begin to be excessive these *two end-products react* with each other to form lactic acid in accordance with the following equation:

$$\underset{\text{(Pyruvic acid)}}{CH_3-\overset{\overset{\displaystyle O}{\|}}{C}-COOH} + NADH + H^+ \xrightarrow[\text{dehydrogenase}]{\text{lactic}}$$

$$\underset{\text{(Lactic acid)}}{CH_3-\overset{\overset{\displaystyle OH}{|}}{\underset{\underset{\displaystyle H}{|}}{C}}-COOH} + NAD^+$$

Thus, under anaerobic conditions, by far the major proportion of the pyruvic acid is converted into lactic acid, which diffuses readily out of the cells into the extracellular fluids and even into the intracellular fluids of other less active cells Therefore, lactic acid represents a type of "sinkhole" into which the glycolytic end-products can disappear, thus allowing glycolysis to proceed far longer than would be possible if the pyruvic acid and hydrogen were not removed from the reacting medium. Indeed, glycolysis could proceed for only a few seconds without this conversion Instead, it can proceed for several minutes, supplying the body with considerable quantities of ATP even in the absence of respiratory oxygen.

Reconversion of Lactic Acid and Pyruvic Acid to Glucose in the Presence of Oxygen. The entire schema of glycolysis, all the way from glucose to lactic acid, is theoretically reversible, but in practice the reverse change of pyruvic acid into phosphoenolpyruvic acid is extremely slow. However, other enzymes cause the pyruvic acid to be converted first into oxaloacetic acid and then into phosphoenolpyruvic acid, after which the scheme of reconversion can proceed. Unfortunately, however, this reverse process consumes 6 ATP molecules in comparison with the 2 ATP molecules formed during glycolysis.

When a person begins to breathe oxygen again after a period of anaerobic metabolism, the extra NADH and H^+ as well as the extra pyruvic acid that have built up in the body fluids are rapidly oxidized, thereby greatly reducing their concentrations. As a result, the chemical reaction for formation of lactic acid immediately reverses itself, the lactic acid once again becoming pyruvic acid. Large portions of this are immediately utilized by the citric acid cycle to provide additional oxidative energy, and large quantities of ATP are formed. This excess ATP then causes as much as three-fourths of the remaining excess pyruvic acid to be converted back into glucose.

Thus, the great amount of lactic acid that forms during anaerobic glycolysis does not become lost from the body, for when oxygen is again available, the lactic acid either can be reconverted to glucose or can be [illegible] energy. By far the greater proportion

sion occurs in the liver, but a small amount can also occur in other tissues.

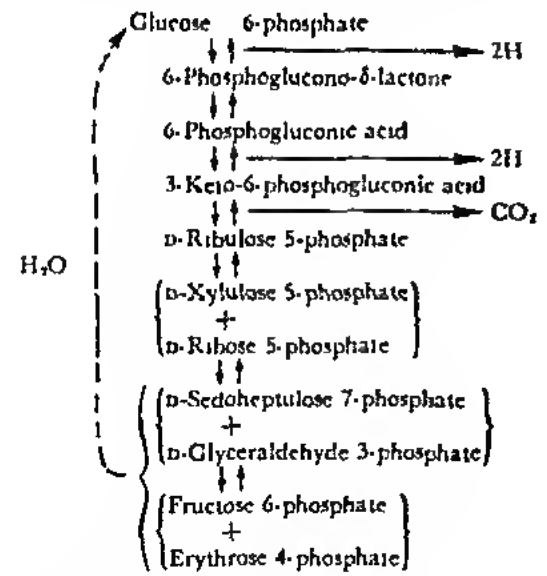
Use of Lactic Acid by the Heart for Energy. Heart muscle is especially capable of converting lactic acid to pyruvic acid and then utilizing this for energy. This occurs to a great extent in very heavy exercise, during which large amounts of lactic acid are released into the blood from the skeletal muscles.

RELEASE OF ENERGY FROM GLUCOSE BY THE PHOSPHOGLUCONATE PATHWAY

Though, in muscle, essentially all the carbohydrates utilized for energy are degraded to pyruvic acid by glycolysis and then oxidized, the glycolytic schema is not the only means by which glucose can be degraded and then oxidized to provide energy. A second important schema for breakdown and oxidation of glucose is called the *phosphogluconate pathway*, which is responsible for as much as 30 per cent of the glucose breakdown *in the liver and even more than this in fat cells*. It is especially important because it can provide energy independently of all the enzymes of the citric acid cycle and therefore is an alternate pathway for energy metabolism in some enzymatic abnormalities of the cells.

Release of Carbon Dioxide and Hydrogen by Means of the Phosphogluconate Pathway. Figure 67–6 illustrates most of the basic chemical reactions in the phosphogluconate pathway. This shows that glucose, after several stages of conversion, releases 1 molecule of carbon dioxide and 4 atoms of hydrogen, with resultant formation of a five-carbon sugar, D-ribulose. This substance in turn can change progressively into several other five-, four-, seven-, and three-carbon sugars. Finally, various combinations of these sugars can resynthesize glucose. However, *only 6 molecules of glucose are resynthesized for every 6 molecules of glucose that initially enter into the reactions*. That is, the phosphogluconate pathway is a cyclic process in which 1 molecule of glucose is metabolized for each revolution of the cycle. Thus, by revolution of the cycle again and again, all the glucose can eventually be converted into carbon dioxide and hydrogen, and the hydrogen in turn can enter the oxidative phosphorylation pathway to form ATP, or more often it is used for synthesis of fat as follows.

Use of Hydrogen to Synthesize Fat; The Function of Nicotinamide Adenine Dinucleotide Phosphate ($NADP^+$). The hydrogen released during the phosphogluconate cycle does not combine with NAD^+ as in the glycolytic pathway, but combines with $NADP^+$, which is almost identical with NAD^+ except for an extra phosphate radical. Yet, this difference is extremely significant because only hydrogen bound with $NADP^+$ in the form of NADPH can be used for synthesis of fat from carbohydrates, which is discussed in the following chapter. When glycolysis becomes slowed because of cellular inactivity, the phosphogluconate pathway still remains operative (mainly in the liver) to break down any excess glucose that continues to be transported into the cells, and NADPH becomes abundant to help convert acetyl radicals into long fatty acid chains.



Net reaction

$$\text{Glucose} + 12NADP^+ + 6H_2O \longrightarrow 6CO_2 + 12H + 12NADPH$$

Figure 67–6. The phosphogluconate pathway for glucose metabolism.

GLUCOSE CONVERSION TO GLYCOGEN OR FAT

When glucose is not immediately required for energy, the extra glucose that continually enters the cells either is stored as glycogen or is converted into fat. Glucose is preferentially stored as glycogen until the cells have stored as much glycogen as they can—an amount sufficient to supply the energy needs of the body for only 1 to 24 hours. When the cells (primarily liver and muscle cells) approach saturation with glycogen, the additional glucose is then converted into fat in the liver and in fat cells and then is stored in the fat cells. Other steps in the chemistry of this conversion are discussed in the following chapter.

FORMATION OF CARBOHYDRATES FROM PROTEINS AND FATS—"GLUCONEOGENESIS"

When the body's stores of carbohydrates decrease below normal, moderate quantities of glucose can be formed from *amino acids* and from the *glycerol* portion of fat. This process is called *gluconeogenesis*. Approximately 60 per cent of the amino acids in the body proteins can be converted easily into carbohydrates, while the remaining 40 per cent have chemical configurations that make this difficult. Each amino acid is converted into glucose by a slightly different chemical process. For instance, alanine can be converted directly into pyruvic acid simply by deamination, the pyruvic acid then is converted into glucose, as explained previously. Several of the more complicated amino acids can be converted into different sugars containing three-, four-, five-, or seven-carbon atoms; these can then enter the phosphogluconate pathway and eventually form glucose. Thus, by means of deamination plus several simple interconversions, many of the amino acids become glucose. Similar interconversions can change glycerol into carbohydrates.

Regulation of Gluconeogenesis. Diminished carbohydrates in the cells and decreased blood sugar are the basic stimuli that set off an increase in the rate of gluconeogenesis. The diminished carbohydrates can directly cause reversal of many of the glycolytic and phosphoglucanate reactions, thus allowing conversion of deaminated amino acids and glycerol into carbohydrates. However, in addition, several of the hormones secreted by the endocrine glands are especially important in this regulation, as follows:

Effect of Corticotropin and Glucocorticoids on Gluconeogenesis. When normal quantities of carbohydrates are not available to the cells, the adenohypophysis, for reasons not completely understood at present, begins to secrete increased quantities of *corticotropin*, which stimulate the adrenal cortex to produce large quantities of *glucocorticoid hormones*, especially *cortisol*. In turn, cortisol mobilizes proteins from essentially all cells of the body, making these available in the form of amino acids in the body fluids. A high proportion of these immediately become deaminated in the liver and therefore provide ideal substrates for conversion into glucose. Thus, one of the most important means by which gluconeogenesis is promoted is through the release of glucocorticoids from the adrenal cortex.

Effect of Thyroxine on Gluconeogenesis. Thyroxine, secreted by the thyroid gland, also increases the rate of gluconeogenesis. This, too, is believed to result principally from mobilization of proteins from the cells. However, it might result to some extent also from the mobilization of fats from the fat depots, the glycerol portion of the fats being converted into glucose.

BLOOD GLUCOSE

The normal blood glucose concentration in a person who has not eaten a meal within the past three to four hours is approximately 90 mg/dl and, even after a meal containing large amounts of carbohydrates, this rarely rises above 140 mg/dl unless the person has diabetes mellitus, which will be discussed in Chapter 78.

The regulation of blood glucose concentration is intimately related to the hormones insulin and glucagon, this subject will be discussed in detail in Chapter 78 in relation to the functions of these hormones.

REFERENCES

Aoki, T R., and Cahill, G F , Jr Metabolic effects of insulin, glucagon, and glucose in man. Clinical applications. *In* DeGroot, L J , *et al.* (eds) Endocrinology. Vol 3 New York, Grune & Stratton, 1979, p 1843.

Baer, H P , and Drummond, G. I. (eds) Physiological and Regulatory Functions of Adenosine and Adenine Nucleotides New York, Raven Press, 1979

Baltscheffsky, H , and Baltscheffsky, M , Electron transport phosphorylation *Annu. Rev Biochem , 43* 871, 1974

Butler, T. M , and Davies, R E High-energy phosphates in smooth muscle *In* Bohr, D F , *et al* (eds) Handbook of Physiology Sec 2, Vol II. Baltimore, Williams & Wilkins, 1980, p 237

Caren, G , *et al.*, Enzyme dynamics The statistical physics approach. *Annu Rev Biophys. Bioeng 8* 69, 1979

Cornish-Bowden. A Fundamentals of Enzyme Kinetics Boston, Butterworths, 1979

Desnick, R. J., *et al* Toward enzyme therapy for lysosomal storage diseases. *Physiol Rev., 58* 57, 1978.

Dickerson, R E Cytochrome C and the evolution of energy metabolism. *Sci Am , 242*(3) 136, 1980

Esmann, V (ed) Regulatory Mechanisms of Carbohydrate Metabolism New York, Pergamon Press, 1978

Felig, P Disorders of carbohydrate metabolism *In* Bondy, P K , and Rosenberg, L. E. (eds) Metabolic Control and Disease, 8th Ed. Philadelphia, W B Saunders Co , 1980, p 276

Foster, R. L The Nature of Enzymology New York, John Wiley & Sons, 1979

Friedmann, H C (ed.) Enzymes. Stroudsburg, Pa. Dowden, Hutchinson & Ross, 1980

Frohman, L. A. CNS peptides and glucoregulation *Annu Rev Physiol , 45* 95, 1983

Galbraith, P D Metabolism of substrates Energy substrate metabolism during exercise and as modified by training *Fed. Proc , 44* 333, 1985

Hasselberger, F X Uses of Enzymes and Immobilized Enzymes Chicago Nelson Hall, 1978.

Hems, D A., and Whatton, P D Control of hepatic glycogenolysis *Physiol Rev , 60*:1, 1980

Hetenyi, G Jr , *et al* Turnover and precursor-product relationships of nonlipid metabolites *Physiol Rev , 63* 606, 1983

Homsher E , and Kean, C J Skeletal muscle energetics and metabolism. *Annu Rev Physiol , 40* 93, 1978

Hoolaway, M R The Mechanism of Enzyme Action. London, Oxford University Press, 1976

Howell, S. L., and Tyhurst, M Insulin secretion. The effector system *Experimentia, 40* 1098, 1984

Huijing, F Glycogen metabolism and glycogen-storage diseases. *Physiol Rev , 55* 609, 1975

Jacquez, J A Red blood cell as glucose carrier Significance for placental and cerebral glucose transfer *Am J Physiol , 246* R289, 1984

Jeffery, J (ed.) Dehydrogenases Requiring Nicotinamide Coenzymes Boston, Birkhauser, 1979

Klachko, D M, *et al* (eds) Hormones and Energy Metabolism New York, Plenum Press, 1978

Kraus-Friedmann, N Hormonal regulation of hepatic gluconeogenesis *Physiol Rev , 64* 170, 1984

Krebs, H A The tricarboxylic and cycle *Harvey Lectures, 44* 165, 1948-1949

Lund-Andersen, H. Transport of glucose from blood to brain *Physiol Rev , 549* 305, 1979

Martin, D B Metabolism and energy mechanisms. *In* Frohlich, E. D (ed) Pathophysiology, 2nd Ed Philadelphia, J B Lippincott Co , 1976, p 365

McCarty, R. E.. How cells make ATP *Sci Am , 238*(3) 104, 1978.

McGilvery, R. W Biochemistry A Functional Approach, 2nd Ed Philadelphia, W B Saunders Co , 1979

Oomura, Y, and Yoshimatsu, H Neural network of glucose monitoring system. *J Auton Nerv Syst , 10*:359, 1984

Purich, D L (ed) Enzyme Kinetics and Mechanism. New York, Academic Press, 1979

Racker, E (ed) Energy Transducing Mechanisms. Baltimore, University Park Press, 1975

Singer, T P , and Dndarza, P N (eds) Mechanisms of Oxidizing Enzymes New York, Elsevier-North-Holland, 1978

Storlien, L. H The role of the ventromedial hypothalamic area in peripancreatic glucoregulation *Life-Sci , 36* 505, 1985

Wang, J H Coupling of proton flux to the hydrolysis and synthesis of ATP *Annu Rev Biophys Bioeng , 12* 21, 1983

Wynn, C H The Structure and Function of Enzymes, 2nd Ed. Baltimore, University Park Press, 1979

68

ipid Metabolism

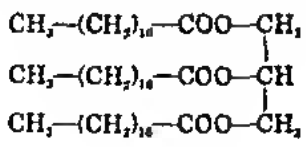
A number of different chemical compounds in the food and in the body are classified as *lipids* These include (1) *neutral fat*, known also as *triglycerides*, (2) the *phospholipids*, (3) *cholesterol*, and (4) a few others of less importance. Chemically, the basic lipid moiety of both the triglycerides and the phospholipids is *fatty acids*, which are simply long chain hydrocarbon organic acids A typical fatty acid, palmitic acid, is the following

$$CH_3(CH_2)_{14}COOH$$

Though cholesterol does not contain fatty acid, its sterol nucleus, as pointed out later in the chapter, is synthesized from degrodation products of fatty acid roolecules, thus giving it many of the physical and chemical properties of other lipid substances

Tha triglycerides are used in the body roainly to provide energy for the different metabolic processes; this function thay shara almost equally with the carbohydrates However, some lipids, especially cholesterol, the phospholipids, and derivatives of these, are used throughout the body to provide other intracellular functions.

Basic Chemical Structure of Triglycerides (Neutral Fat). Since most of this chapter deals with utilization of triglycerides for energy, the following basic stcucture of the triglyceride molecule must be understood.



Tristearin

Note that three long chain fatty acid molecules are bound with one molecule of glycerol. In the human body, the three fatty acids most commonly present in neutral fat are (1) *stearic ocid*, which has an 18-carbon chain and is fully saturated with hydrogen atoms, (2) *oleic acid*, which also has an 18-carbon chain but has one double bond in the middle of the chain, and (3) *palmitic acid*, which has 16 carbon atoms and is fully saturated.

TRANSPORT OF LIPIDS IN THE BLOOD

TRANSPORT FROM THE GASTROINTESTINAL TRACT—THE CHYLOMICRONS

It will be recalled from Chapter 65 that almost all the fats of the diet are absorbed into the lymph In the digestive tract, most triglycerides are split into monoglycerides and fatty acids Then, on passing through the intestinal epithelial cells, these are resynthesized into new molecules of triglycerides which aggregate and enter the lymph as minute, dispersed droplets called *chylomicrons*, having sizes between 0.03 and 0 5 micron. A small amount of the protein apoprotein B adsorbs to the outer surfaces of the chylomicrons; this increases their suspension atability in the fluid of the lymph and prevents their adherence to the lymphatic vessel walls.

Most of the cholesterol and phospholipids absorbed from the gastrointestinal tract, as well as sroall amounts of phospholipids that are continually synthesized by the intestinal mucosa, also enter the chylomicrons Thus, chyloroicrons are composed principally of triglycerides, but they also contain approximately 9 per cent phospholipids, 3 por cent cholesterol, and 1 per cent apoprotein B as well The chylomicrons are then transported up the thoracic duct and emptied into the venous blood at the juncture of the jugular and subclavian veins.

Rerooval of the Chylomicrons from the Blood. An hour or so after a meol that contsins large quantities of fat, the chylomicron concentration in the plasma may rise to as high as 1 to 2 per cent, and, because of the large sizes of the chylomicrons, the plasma appears turbid and sometimes yellow However, the chylomicrons have a half-life of less than an hour so that the plasma becomes clear again within a few hours The fat of the chylomicrons is removed mainly in the following way.

Hydrolysis of the Chylomicron Triglycerides by Lipoprotein Lipase; Fat Storage in the Fat and Liver Cells. Most of the chylomicrons are removed from the circulating blood as they pass through the capillaries of adipose tissue and the liver. Both adipose tissue and the liver contain large quantities of the enzyme called *lipoprotein lipase* This enzyme is active in the capillary endothelium where it hydrolyzes the triglycerides of chylomicrons that stick to the endothelial wall, releasing fatty acids and glycerol The fatty acids, being highly

miscible with the membranes of the cells, immediately diffuse into the fat and liver cells. Once within these cells, the fatty acids are resynthesized into triglycerides, new glycerol being supplied by the metabolic processes of the cells, as will be discussed later in the chapter. The lipase also causes hydrolysis of phospholipids, this too releasing fatty acids to be stored in the cells in the same way. Thus, essentially all the mass of the chylomicrons is removed from the circulating blood.

TRANSPORT OF FATTY ACIDS IN COMBINATION WITH ALBUMIN—"FREE FATTY ACID"

When the fat that has been stored in the fat cells is to be used elsewhere in the body, usually to provide energy, it must first be transported to the other tissues. It is transported almost entirely in the form of *free fatty acid*. This is achieved by hydrolysis of the triglycerides once again into fatty acids and glycerol At least two classes of stimuli play important roles in promoting this hydrolysis. First, when the availability of glucose to the cell falls too low, one of its breakdown products, *α-glycerophosphate*, also becomes too low This substance is required to form the glycerol portion of newly synthesized triglycerides, and in its absence the equilibrium shifts in favor of hydrolysis. Second, a *hormone-sensitive cellular lipase* can be activated by a multitude of different hormones, and this promotes rapid hydrolysis of the triglyceride. This also will be discussed later in the chapter.

On leaving the fat cells, the fatty acids ionize strongly in the plasma and immediately combine with albumin of the plasma proteins. The fatty acid bound with proteins in this manner is called *free fatty acid* or *nonesterified fatty acid* (or simply *FFA* or *NEFA*) to distinguish it from other fatty acids in the plasma that exist in the form of esters of glycerol, cholesterol, or other substances.

The concentration of free fatty acid in the plasma under resting conditions is about 15 mg/dl, which is a total of only 0.45 gram of fatty acids in the entire circulatory system. Yet, strangely enough, even this small amount accounts for almost all the transport of fatty acids from one part of the body to another for the following reasons:

(1) Despite the minute amount of free fatty acid in the blood, its rate of "turnover" is extremely rapid; *half the plasma fatty acid is replaced by new fatty acid every two to three minutes*. One can calculate that at this rate almost all the normal energy requirements of the body can be provided by oxidation of the transported free fatty acid without using any carbohydrates or proteins for energy. (2) All conditions that increase the rate of utilization of fat for cellular energy also increase the free fatty acid concentration in the blood; this sometimes increases as much as five- to eightfold. Especially does this occur in starvation and in diabetes, in both of which conditions the person derives little or no energy from carbohydrates.

Under normal conditions, about 3 molecules of fatty acid combine with each molecule of albumin, but as many as 30 fatty acid molecules can combine with a single albumin molecule when the need for fatty acid transport is extreme. This shows how variable the rate of lipid transport can be under different physiological needs.

THE LIPOPROTEINS

In the postabsorptive stote—that is, after all the chylomicrons have been removed from the blood—over 95 per cent of all the lipids in the plasma are in the form of lipoproteins, which are small particles much smaller than chylomicrons but qualitatively similar in composition, containing mixtures of *triglycerides, phospholipids, cholesterol*, and *protein*. The protein in the mixture averages about one fourth to one third of the total constituents, and the remainder is lipids The total concentration of lipoproteins in the plasma averages about 700 mg/dl, and this can be broken down into the following average concentrations of the individual constituents:

	mg/dl of plasma
Cholesterol	180
Phospholipids	160
Triglycerides	160
Protein	200

Types of Lipoproteins. Chylomicrons are sometimes classified as lipoproteins because they contain both lipids and protein. In addition to the chylomicrons, however, there are three other major classes of lipoprotein based on their densities, as measured in the ultracentrifuge: (1) *very low density lipoproteins*, which contain high concentrations of triglycerides and moderate concentrations of both phospholipids and cholesterol; (2) *low density lipoproteins*, which contain relatively little triglyceride but a very high percentage of cholesterol; and (3) *high density lipoproteins*, which contain about 50 per cent protein with smaller concentrations of the lipids.

Formation of the Lipoproteins. The lipoproteins are formed almost entirely in the liver, which is in keeping with the fact that most plasma phospholipids, cholesterol, and triglycerides (except those in the chylomicrons) are synthesized in the liver. However, small quantities of high density lipoproteins are synthesized in the intestinal epithelium during the absorption of fatty acids from the intestines

Function of the Lipoproteins. The principal function of the lipoproteins in the plasma is to transport lipids throughout the body For instance, the turnover of triglycerides in the lipoproteins is as much as several grams per hour and perhaps half this much turnover of cholesterol and phospholipids.

Triglycerides are synthesized mainly from carbohydrates in the liver and are transported to the adipose tissue and other peripheral tissues in the *very low density lipoproteins*. The *low density lipoproteins* are the residuals of the very low density lipoproteins after they have delivered most of their triglycerides to the adipose tissue, leaving large concentrations of cholesterol and phospholipids in the low density lipoproteins. On the other hand, the *high density lipoproteins* transport cholesterol away from the peripheral tissues and to the liver; therefore, this type of lipoprotein probably plays a very important role in preventing the development of atherosclerosis, as we shall discuss later in the chapter.

THE FAT DEPOSITS

ADIPOSE TISSUE

Large quantities of fat are frequently stored in two major tissues of the body, the adipose tissue and the liver. The adipose tissue is usually called the *fat deposits*, or simply the *fat depots*

The major function of adipose tissue is storage of triglycerides until these are needed to provide energy elsewhere in the body However, a subsidiary function is to provide heat insulation for the body, as will be discussed in Chapter 72.

The Fat Cells. The fat cells of adipose tissue are modified fibroblasts that are capable of storing almost pure triglycerides in quantities equal to 80 to 95 per cent of their volume. The triglycerides are generally in a liquid form, and when the tissues of the skin are exposed to prolonged cold, the fatty acid chains of the triglycerides, over a period of weeks, become either shorter or more unsaturated to decrease their melting point, thereby always allowing the fat in the fat cells to remain in a liquid state. This is particularly important because only liquid fat can be hydrolyzed and then transported from the cells

Fat cells can synthesize small amounts of fatty acids and triglycerides from carbohydrates, this function supplementing the synthesis of fat in the liver, as discussed later in the chapter

Exchange of Fat Between the Adipose Tissue and the Blood—Tissue Lipases. As discussed above, large quantities of lipases are present in adipose tissue Some of these enzymes catalyze the deposition of triglycerides from the chylomicrons and other lipoproteins Others, when activated by hormones, cause splitting of the triglycerides of the fat cells to release free fatty acids Because of rapid exchanges of the fatty acids, the triglycerides in the fat cells are renewed approximately once every two to three weeks, which means that the fat stored in the tissues today is not the same fat that was stored last month, thus emphasizing the dynamic state of the storage fat.

THE LIVER LIPIDS

The principal functions of the liver in lipid metabolism are (1) to degrade fatty acids into small compounds that can be used for energy, (2) to synthesize triglycerides, mainly from carbohydrates but to a lesser extent from proteins as well, and (3) to synthesize other lipids from fatty acids, especially cholesterol and phospholipids

Large quantities of triglycerides appear in the liver (a) during starvation, (b) in diabetes mellitus, or (c) in any other condition in which fat is being utilized rapidly for energy In these conditions, large quantities of triglycerides are mobilized from the adipose tissue, transported as free fatty acids in the blood, and then redeposited as triglycerides in the liver, where the initial stages of much of the fat degradation begin. Thus, under normal physiological conditions the total amount of triglycerides in the liver is controlled to a great extent by the overall rate at which lipids are being utilized for energy

The liver cells, in addition to containing triglycerides, contain large quantities of phospholipids and cholesterol, which are continually synthesized by the liver. Also, the liver cells are much more capable than other tissues of desaturating fatty acids so that the liver triglycerides normally are much more unsaturated than the triglycerides of the adipose tissue This capability of the liver to desaturate fatty acids is functionally important to all the tissues of the body, because many of the structural members of all cells contain reasonable quantities of unsaturated fats, and their principal source is the liver This desaturation is accomplished by a dehydrogenase in the liver cells

USE OF TRIGLYCERIDES FOR ENERGY, AND FORMATION OF ADENOSINE TRIPHOSPHATE (ATP)

Approximately 40 to 45 per cent of the calories in the normal American diet are derived from fats, which is about equal to the calories derived from carbohydrates Therefore, the use of fats by the body for energy is equally as important as the use of carbohydrates. In addition, much of the carbohydrates ingested with each meal is converted into triglycerides, then stored, and later utilized as triglycerides for energy.

Hydrolysis of the Triglycerides. The first stage in the utilization of triglycerides for energy is hydrolysis of these compounds into fatty acids and glycerol and subsequent transport of both these products to the active tissues where they will be oxidized to give energy Almost all cells, with the notable exception of brain tissue, can use fatty acids almost interchangeably with glucose for energy

The glycerol, on entering the active tissue, is immediately changed by intracellular enzymes into glycero 3-phosphate, which enters the glycolytic pathway for glucose breakdown and in this way is used for energy However, before the fatty acids can be used for energy they must be processed further in the following way:

Entry of Fatty Acids into the Mitochondria. The degradation and oxidation of fatty acids occur only in the mitochondria Therefore, the first step for utilization of the fatty acids is their transport into the mitochondria This is an enzyme-catalyzed process that employs *carnitine* as a carrier substance. Once inside the mitochondria, the fatty acid splits away from the carnitine and is then oxidized

Degradation of Fatty Acid to Acetylcoenzyme A by Beta Oxidation. The fatty acid molecule is degraded in the mitochondria by progressive release of 2-carbon segments in the form of acetylcoenzyme A (acetyl-CoA) This process, which is illustrated in Figure 68–1, is called the *beta oxidation* process for degradation of fatty acids Let us summarize the results of the equations in this figure

1. The beta carbon of the fatty acid molecule is oxidized by adding a water molecule, then removing two hydrogen ions, leaving the oxygen from the water bound with the beta carbon.
2. The molecule then splits between the alpha and beta carbons, releasing 1 molecule of acetyl-CoA Thus, the fatty acid molecule becomes two carbon atoms shorter.
3. Then the process repeats itself again and again until the entire fatty acid chain has been split into multiple acetyl-CoA molecules
4. For each acetyl-CoA molecule split from the fatty acid, a total of four hydrogen atoms is released. These

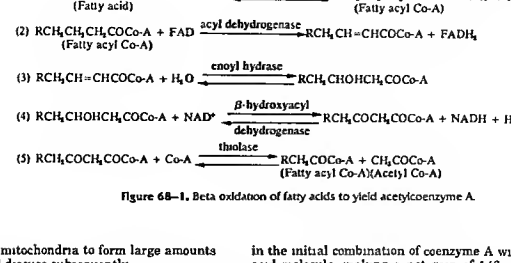


Figure 68–1. Beta oxidation of fatty acids to yield acetylcoenzyme A.

are oxidized in the mitochondria to form large amounts of ATP, as we shall discuss subsequently.

Oxidation of Acetyl-CoA. The acetyl-CoA molecules formed by beta oxidation of fatty acids enter into the citric acid cycle as explained in the preceding chapter, combining first with oxaloacetic acid to form citric acid, which then is degraded into carbon dioxide and hydrogen atoms. The hydrogen is subsequently oxidized by the oxidative system of the mitochondria, which was explained in the preceding chapter. The net reaction for each molecule of acetyl-CoA is the following:

$$CH_3COCo\text{-}A + \text{Oxaloacetic acid} + 3H_2O + ADP$$

Citric acid cycle

$$2CO_2 + 8H + HCo\text{-}A + ATP + \text{Oxaloacetic acid}$$

Thus, after the initial degradation of fatty acids to acetyl-CoA, their final breakdown is precisely the same as that of the acetyl-CoA formed from pyruvic acid during the metabolism of glucose

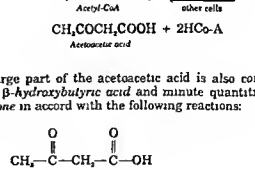
ATP Formed by Oxidation of Fatty Acid. In Figure 68–1 note also that 4 hydrogen atoms are released each time a molecule of acetyl-CoA is formed from the fatty acid chain. Therefore, for every stearic acid molecule that is split, a total of 32 hydrogen atoms is removed In addition, for each acetyl-CoA degraded by the citric acid cycle, 8 hydrogen atoms are removed, making an additional 72 hydrogens for each molecule of stearic acid metabolized. This, added to the above 32 hydrogen atoms, makes a total of 104 hydrogen atoms Of this group 34 are removed from the degrading fatty acid by flavoproteins and 70 are removed by NAD^+ as NADH and H^+. These two groups of hydrogen atoms are oxidized in the mitochondria, as discussed in the preceding chapter, but they enter the oxidative system at different points, so that up to 1 molecule of ATP is synthesized for each of the 34 flavoprotein hydrogens and up to 1.5 molecules of ATP are synthesized for each of the 70 NADH and H^+ hydrogens. This makes 34 plus 105, or a total of 139 molecules of ATP formed by the oxidation of hydrogen derived from each molecule of stearic acid. And another 9 molecules of ATP are formed in the citric acid cycle, one for each of the 9 acetyl-CoA molecules metabolized. Thus, a total of 148 molecules of ATP is formed during the complete oxidation of 1 molecule of stearic acid. However, 2 high energy bonds are consumed in the initial combination of coenzyme A with the fatty acid molecule, making a net gain of 146 molecules of ATP.

FORMATION OF ACETOACETIC ACID IN THE LIVER AND ITS TRANSPORT IN THE BLOOD

A large share of the initial degradation of fatty acids occurs in the liver, especially when excessive amounts of lipids are being mobilized for energy However, the liver uses only a small proportion of the fatty acids for its own intrinsic metabolic processes Instead, when the fatty acid chains have been split into acetyl-CoA, 2 molecules of acetyl-CoA condense to form 1 molecule of acetoacetic acid that is then transported in the blood to the other cells throughout the body where it is used for energy The chemical processes are the following:

$$\underset{\text{Acetyl-CoA}}{2CH_3COCo\text{-}A} + H_2O \underset{\text{other cells}}{\overset{\text{liver cells}}{\rightleftharpoons}} \underset{\text{Acetoacetic acid}}{CH_3COCH_2COOH} + 2HCo\text{-}A$$

A large part of the acetoacetic acid is also converted into β-*hydroxybutyric acid* and minute quantities into *acetone* in accord with the following reactions:



The acetoacetic acid and β-hydroxybutyric acid diffuse freely through the liver cell membranes and are transported by the blood to the peripheral tissues. Here they again diffuse into the cells where reverse reactions occur and acetyl-CoA molecules are formed. These in turn enter the citric acid cycle and are oxidized for energy, as explained.

Normally, the acetoacetic acid and β-hydroxybutyric acid that enter the blood are transported so rapidly to the tissues that their combined concentration in the plasma rarely rises above 3 mg/dl. Yet despite the small quantities in the blood, large amounts are actually transported; this is analogous to the high rate of free fatty acid transport. The rapid transport of both these substances depends on their high degree of lipid solubility, which allows rapid diffusion through the cell membranes

Ketosis and Its Occurrence in Starvation, Diabetes, and Other Diseases. Large quantities of acetoacetic acid, β-hydroxybutyric acid, and acetone occasionally accumulate in the blood and interstitial fluids; this condition is called *ketosis* because acetoacetic acid is a keto acid, and the three compounds are called *ketone bodies* Ketosis occurs especially in starvation, in diabetes mellitus, or sometimes even when a person's diet is composed almost entirely of fat. In all these states, essentially no carbohydrates are metabolized—in starvation and following a high fat diet because carbohydrates are not available and in diabetes because insulin is not available to cause glucose transport into the cells.

When carbohydrates are not utilized for energy, almost all the energy of the body must come from metabolism of fats We shall see later in the chapter that lack of availability of carbohydrates automatically increases the rate of removal of fatty acids from adipose tissues, and in addition, several hormonal factors—such as increased secretion of corticotropin by the adenohypophysis, increased secretion of glucocorticoids by the adrenal cortex, and decreased secretion of insulin by the pancreas—all further enhance the removal of fatty acids from the fat tissues As a result, tremendous quantities of fatty acids become available to the liver for degradation. The ketone bodies in turn pour out of the liver to be carried to the cells. Yet the cells are limited in the amount of ketone bodies that can be oxidized for several reasons, the most important of which is the following One of the products of carbohydrate metabolism is the oxaloacetate that is required to bind with acetyl-CoA before it can be processed in the citric acid cycle Therefore, deficiency of oxaloacetate limits the entry of acetyl-CoA into the citric acid cycle, and because of the simultaneous outpouring of tremendous quantities of acetoacetic acid and the other ketone bodies from the liver, the blood concentration of acetoacetic acid and β-hydroxybutyric acid sometimes rises to as high as 20 to 30 times normal, thus leading to extreme acidosis, as explained in Chapter 37.

The acetone that is formed during ketosis is a volatile substance that is blown off in small quantities in the expired air of the lungs, often giving the breath an acetone smell. This smell is frequently used as a diagnostic criterion of ketosis.

Adaptation to a High Fat Diet. Upon changing slowly from a carbohydrate diet to an almost completely fat diet, a person's body adapts to the utilization of far more acetoacetic acid than usual, and, in this instance ketosis does not occur. For instance, the Eskimos, who sometimes live almost entirely on a fat diet, do not develop ketosis. Presumably some factor enhances the rate of acetoacetic acid metabolism by the cells.

SYNTHESIS OF TRIGLYCERIDES FROM CARBOHYDRATES

Whenever a greater quantity of carbohydrates enters the body than can be used immediately for energy or stored in the form of glycogen, the excess is rapidly converted into triglycerides and then stored in this form in the adipose tissue In human beings, most triglyceride synthesis occurs in the liver, but small quantities are also synthesized in the adipose tissue. The triglycerides formed in the liver are transported by the very low density lipoproteins to the adipose tissue.

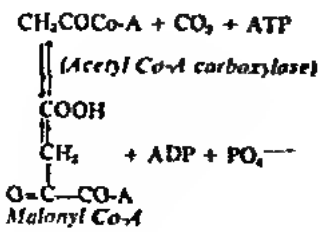
Conversion of Acetyl-CoA into Fatty Acids. The first step in the synthesis of triglycerides is conversion of carbohydrates into acetyl-CoA. It will be recalled from the preceding chapter that this occurs during the normal degradation of glucose by the glycolytic system It will also be remembered from earlier in this chapter that fatty acids are actually large polymers of acetic acid. Therefore, it is easy to understand how acetyl-CoA could be converted into fatty acids

However, synthesis of by far the largest proportion of the fatty acids from acetyl-CoA is not achieved by simply reversing the oxidative degradation that was described above Instead, this occurs by the two-step process shown in Figure 68–2, using *malonyl CoA* and NADPH as the principal intermediates in the polymerization process.

Combination of Fatty Acids with α-Glycerophosphate to Form Triglycerides. Once the synthesized fatty acid chains have grown to contain 14 to 18 carbon atoms, they then become bound to glycerol to form triglycerides The enzymes that cause this conversion are highly specific for fatty acids with chain lengths of 14 carbon atoms or greater, a factor that actually controls the physical quality of the triglycerides stored in the body.

As illustrated in Figure 68–3, the glycerol portion of

Step 1:



Step 2:

1 Acetyl Co-A + 8 Malonyl Co-A + 16NADPH + 16H⁺ ———→

1 Stearic Acid + 8CO₂ + 9Co-A + 16NADP⁺ + 7H₂O

Figure 68–2. Synthesis of fatty acids.

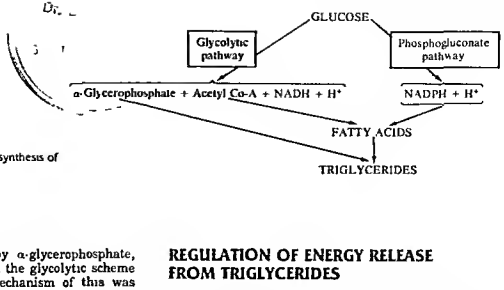


Figure 68–3. An overall schema for synthesis of triglycerides from glucose.

the triglyceride is furnished by α-glycerophosphate, which is a product derived from the glycolytic scheme of glucose degradation. The mechanism of this was discussed in Chapter 67.

Efficiency of Carbohydrate Conversion into Fat. During triglyceride synthesis, only about 15 per cent of the original energy in the glucose is lost in the form of heat, while the remaining 85 per cent is transferred to the stored triglycerides.

Importance of Fat Synthesis and Storage. Fat synthesis from carbohydrates is especially important for two reasons: (1) The ability of the different cells of the body to store carbohydrates in the form of glycogen is generally slight; a maximum of only a few hundred grams of glycogen can be stored in the liver, the skeletal muscles, and all other tissues of the body put together. In contrast, many kilograms of fat can be stored. Therefore, fat synthesis provides a means by which the energy of excess ingested carbohydrates (and proteins, too) can be stored for later use. Indeed, the average person has almost 150 times as much energy stored in the form of fat as stored in the form of carbohydrate. (2) Each gram of fat contains 2.25 times as many calories of energy as each gram of glycogen. Therefore, for a given weight gain, a person can store far more energy in the form of fat than in the form of carbohydrate, which is exceedingly important when an animal must be highly motile to survive.

Failure to Synthesize Fats from Carbohydrates in the Absence of Insulin. When insulin is not available, as in diabetes mellitus, fats are poorly if at all synthesized. This results from the following effects. First, when insulin is not available, glucose does not enter the fat and liver cells satisfactorily, so that little of the acetyl-CoA and NADPH needed for fat synthesis can be derived from glucose. Second, lack of glucose in the fat cells greatly reduces the availability of α-glycerophosphate, which also makes it difficult for the tissues to form triglycerides.

SYNTHESIS OF TRIGLYCERIDES FROM PROTEINS

Many amino acids can be converted into acetyl-CoA, as will be discussed in the following chapter. Obviously, this can be synthesized into triglycerides. Therefore, when people have more proteins in their diets than their tissues can use as proteins, a large share of the excess is stored as fat.

REGULATION OF ENERGY RELEASE FROM TRIGLYCERIDES

Fat-Sparing Effect of Carbohydrates; Control of Fat Synthesis from Carbohydrates. When adequate quantities of carbohydrates are available in the body, the utilization of triglycerides for energy is greatly depressed. In place of fat utilization, the carbohydrates are utilized preferentially. There are several different reasons for this "fat-sparing" effect of carbohydrates. One of the most important is the following: The fats in adipose tissue cells are present in two different forms, triglycerides and small quantities of free fatty acids. These are in constant equilibrium with each other. When excess quantities of *α-glycerophosphate* are present, the equilibrium between free fatty acids and triglycerides shifts toward the triglycerides; as a result, only minute quantities of fatty acids are then available to be utilized for energy. Since α-glycerophosphate is an important product of glucose metabolism, the availability of large amounts of glucose automatically inhibits the use of fatty acids for energy.

Second, when carbohydrates are available in excess, fatty acids are synthesized more rapidly than they are degraded. This effect is caused partially by the large quantities of acetyl-CoA formed from the carbohydrates and by the low concentration of free fatty acids in the adipose tissue, thus creating conditions appropriate for conversion of acetyl-CoA into fatty acids. However, an even more important effect that promotes conversion of carbohydrates to fats is the following: The first step, and the rate-limiting step, in the synthesis of fatty acids is carboxylation of acetyl-CoA to form malonyl-CoA, as was discussed above. The rate of this reaction is controlled primarily by the enzyme *acetyl-CoA carboxylase*, the activity of which is accelerated in the presence of the intermediates of the citric acid cycle. When excess carbohydrates are being utilized, these intermediates increase, thus automatically causing increased synthesis of fatty acids. Thus, an excess of carbohydrates in the diet not only acts as a fat-sparer but also increases the fat in the fat stores.

Acceleration of Fat Utilization for Energy in the Absence of Carbohydrates. All the fat-sparing effects of carbohydrates are lost and are actually reversed when carbohydrates are not available or are available in short supply. Therefore, the equilibrium now shifts in the opposite direction, and fat is mobilized from the adipose cells and used for energy in place of the absent carbohydrates.

But also important are several hormonal changes that toke place to promote rapid fatty acid mobilization from adipose tissue. Among the most important of these is a marked decrease in insulin secretion caused by the absence of carbohydrates This not only reduces the rate of glucose utilization by the tissues but also decreases fat synthesis, which further shifts the equilibrium in favor of fat utilization in place of carbohydrates

Hormonal Regulation of Fat Utilization. At least seven of the hormones secreted by the endocrine glands have marked effects on fat utilization. In addition to the important effect of *insulin lack* discussed earlier, some other important hormonal effects on fat metabolism are listed here

Probably the most dramatic increase that occurs in fat utilization is that observed during heavy exercise This results almost entirely from release of *epinephrine* and *norepinephrine* by the adrenal medullae during exercise, as a result of sympathetic stimulation. These two hormones directly activate *hormone-sensitive triglyceride lipase* that is present in abundance in the fat cells, and this causes very rapid breakdown of triglycerides and mobilization of fatty acids Sometimes the free fatty acid concentration in the blood rises as much as eightfold, and the use of these fatty acids for energy is correspondingly increased Other types of stress that activate the sympathetic nervous system can also increase fatty acid mobilization and utilization in a similar manner.

Stress also causes large quantities of *corticotropin* to be released by the anterior pituitary gland, and this, in turn, causes the adrenal cortex to secrete excessive quantities of *glucocorticoids* Both the corticotropin and glucocorticoids activate either the same hormone-sensitive triglyceride lipase as that activated by epinephrine and norepinephrine or a similar lipase. Therefore, this is still another mechanism for increasing the release of fatty acids from fat tissue. When corticotropin and glucocorticoids are secreted in excessive amounts for long periods of time, as occurs in the endocrine disease called Cushing's disease, fats are frequently mobilized to such a great extent that ketosis results Corticotropin and glucocorticoids are then said to have a *ketogenic effect*.

Growth hormone has an effect similar to but less than that of rorticotropin and glucocorticoids in activating the hormone-sensitive lipase. Therefore, growth hormone can also have a mild ketogenic effect.

Finally, *thyroid hormone* causes rapid mobilization of fat, which is believed to result indirectly from an increased overall rate of energy metabolism in all cells of the body under the influence of this hormone. The resulting reduction in acetyl-CoA and other intermediates of both fat and carbohydrate metabolism in the cells would then be a stimulus to cause fat mobilization

The effects of the different hormones on metabolism are discussed further in the chapters dealing with each of them

OBESITY

Obesity means deposition of excess fat in the body. This subject will be discussed in detail in relation to dietary balances in Chapter 73, but briefly it is caused by ingestion of greater amounts of food than can be lized by the body for energy. The excess food, whether , carbohydrates, or proteins, is then stored as fat in the adipose tissue to be used later for energy. Strain rats have been found in which *hereditary obesity* occ In at least one of these strains, the obesity is caused ineffective mobilization of fat from the adipose tis while synthesis and storage of fat continue norma Obviously, such a one-way process causes progres enhancement of the fat stores, resulting in severe ob ity The problem of obesity will be discussed furthe Chapter 73.

PHOSPHOLIPIDS AND CHOLESTEROL

PHOSPHOLIPIDS

The three major types of body phospholipids are *lecithins*, the *cephalins*, and the *sphingomyelins*, typ examples of which are shown in Figure 68–4.

Phospholipids always contain one or more fatty molecules and one phosphoric acid radical, and t

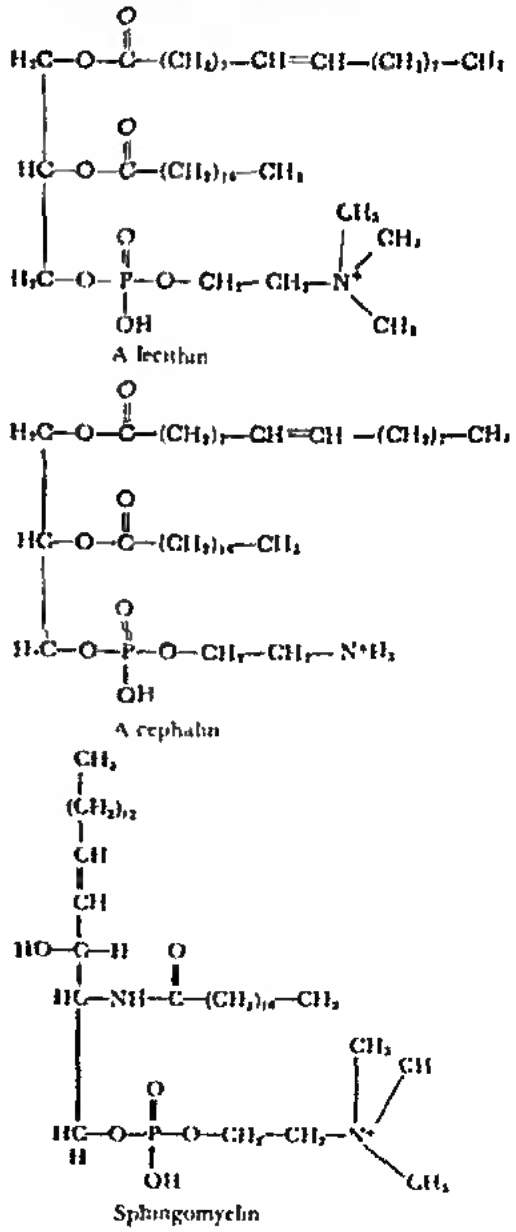


Figure 68–4. Typical phospholipids.

usually contain a nitrogenous base. Though the chemical structures of phospholipids are somewhat variant, their physical properties are similar, for they are all lipid soluble, are transported together in lipoproteins, and are utilized in similar ways throughout the body for various structural purposes.

Formation of Phospholipids. Phospholipids are formed in essentially all cells of the body, though certain cells have a special ability to form them. Probably 90 per cent are formed in the liver cells, though reasonable quantities can also be formed by the intestinal epithelial cells during lipid absorption from the gut.

The rate of phospholipid formation is governed to some extent by the usual factors that control the rate of fat metabolism, for when triglycerides are deposited in the liver, the rate of phospholipid formation increases. Also, certain specific chemical substances are needed for formation of some phospholipids. For instance, *choline*, either in the diet or synthesized in the body, is needed for the formation of lecithin because choline is the nitrogenous base of the lecithin molecule. Also, *inositol* is needed for the formation of some cephalins.

Specific Uses of Phospholipids. Several isolated functions of the phospholipids are the following. (1) Phospholipids are an important constituent of lipoproteins and are essential for the formation of some of these; in their absence serious abnormalities of fat transport can occur. (2) Thromboplastin, which is needed to initiate the clotting process, is composed mainly of one of the cephalins. (3) Large quantities of sphingomyelin are present in the nervous system; this substance acts as an insulator in the myelin sheath around nerve fibers. (4) Phospholipids are donors of phosphate radicals when these are needed for different chemical reactions in the tissues. (5) Perhaps the most important of all the functions of phospholipids is participation in the formation of structural elements—mainly membranes—within the cells throughout the body, as is discussed later in connection with a similar function for cholesterol.

CHOLESTEROL

Cholesterol, the formula of which is illustrated in Figure 68–5, is present in the diet of all persons, and it can be absorbed slowly from the gastrointestinal tract into the intestinal lymph. It is highly fat soluble but only slightly soluble in water, and it is capable of forming esters with fatty acids. Indeed, approximately 70 per cent of the cholesterol in the lipoproteins of the plasma is in the form of cholesterol esters.

Formation of Cholesterol. Besides the cholesterol absorbed each day from the gastrointestinal tract, which is called *exogenous cholesterol*, an even greater quantity

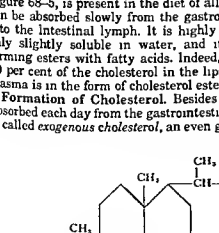


Figure 68–5. Cholesterol.

is also formed in the cells of the body, called *endogenous cholesterol*. Essentially all the endogenous cholesterol that circulates in the lipoproteins of the plasma is formed by the liver, but all the other cells of the body form at least some cholesterol, which is consistent with the fact that many of the membranous structures of all cells are partially composed of this substance.

As illustrated by the formula of cholesterol, its basic structure is a sterol nucleus. This is synthesized entirely from multiple molecules of acetyl-CoA. In turn, the sterol nucleus can be modified by means of various side chains to form (a) cholesterol, (b) cholic acid, which is the basis of the bile acids formed in the liver, and (c) many important steroid hormones secreted by the adrenal cortex, the ovaries, and the testes (these are discussed in later chapters).

Factors That Affect the Plasma Cholesterol Concentration—Feedback Control of Body Cholesterol. Among the important factors that affect plasma cholesterol concentration are the following

1 An increase in the amount of cholesterol ingested each day increases the plasma concentration slightly However, when cholesterol is ingested, the rising concentration of cholesterol inhibits one of the essential enzymes for endogenous synthesis of cholesterol, thus providing an intrinsic feedback control system to regulate plasma cholesterol concentration As a result, plasma cholesterol concentration *usually* is not changed upward or downward more than ±15 per cent by altering the amount of cholesterol in the diet, though *extremes* of cholesterol in the diet can probably alter the level by as much as ±30 per cent

2 A *saturated* fat diet increases blood cholesterol concentration as much as 15 to 25 per cent. This results from increased fat deposition in the liver, which then provides increased quantities of acetyl-CoA in the liver cells for production of cholesterol. Therefore, to decrease the blood cholesterol concentration, it is equally as important to maintain a diet low in saturated fat as to maintain a diet low in cholesterol concentration.

3 Ingestion of fat containing highly unsaturated fatty acids usually depresses the blood cholesterol concentration a slight to moderate amount Unfortunately, the mechanism of this effect is unknown despite the fact that this observation is the basis of much of present-day dietary strategy

4 Lack of thyroid hormone increases the blood cholesterol concentration, whereas excess thyroid hormone decreases the concentration. This effect is believed to be related to the increased metabolism of all lipid substances under the influence of thyroxine.

5 The blood cholesterol also rises greatly in diabetes mellitus. This is believed to result from the general increase in lipid mobilization in this condition.

6 The female sex hormones, the *estrogens*, decrease blood cholesterol, whereas the male sex hormones, the *androgens*, increase blood cholesterol. Unfortunately, the mechanisms of these effects are unknown, but the sex effects are very important because the higher cholesterol in men is associated with a higher incidence of heart attacks at the earlier ages.

Specific Uses of Cholesterol. By far the most abundant use of cholesterol in the body is to form cholic acid in the liver. As much as 80 per cent of the cholesterol is converted into cholic acid. As explained in Chapter 70, this is conjugated with other substances to form bile salts, which promote digestion and absorption of fats, which has already been discussed in connection with fat digestion.

A small quantity of cholesterol is used (a) by the adrenal glands to form *adrenocortical hormones*, (b) the ovaries to form *progesterone* and *estrogen*, and (c) by the testes to form *testosterone* However, these glands can also synthesize their own sterols and then form their hormones from these, as is discussed in the chapters on endocrinology later in the text

A large amount of cholesterol is precipitated in the corneum of the skin This, along with other lipids, makes the skin highly resistant to the absorption of water-soluble substances and also to the action of many chemical agents, for cholesterol and the other lipids are highly inert to such substances as acids and different solvents that might otherwise easily penetrate the body Also, these lipid substances help to prevent water evaporation from the skin, without this protection the amount of evaporation (as occurs in burn patients who have lost their skin) can be as much as 5 to 10 liters per day instead of the usual 300 to 400 milliliters

CELLULAR STRUCTURAL FUNCTIONS OF PHOSPHOLIPIDS AND CHOLESTEROL—ESPECIALLY FOR MEMBRANES

The above-listed uses of phospholipids and cholesterol are of only minor importance in comparison with their importance for general structural purposes throughout the cells of the body, mainly for formation of membranes

In Chapter 2 it was pointed out that large quantities of phospholipids and cholesterol are present in both the cell membrane and the membranes of the internal organelles of all cells It is also known that membrane percentages of both cholesterol and the phospholipids are especially important in determining the fluidity of the cell membranes

For membranes to be formed, substances that are not soluble in water must be available In general, the only substances in the body that are not soluble in water (besides the inorganic substances of bone) are the lipids and some proteins Thus, the physical integrity of cells throughout the body is based mainly on phospholipids, cholesterol, and certain insoluble proteins The polar charges on the phospholipids especially give them the important property of helping decrease the interfacial tension between the membranes and the surrounding fluids

Another fact that indicates the importance of phospholipids and cholesterol for the formation of structural elements of the cells is the slow turnover rate of these substances in most nonhepatic tissues For instance, radioactive phospholipids formed in the brain of mice remain in the brain several months after they are formed Thus, these phospholipids are only slowly metabolized, and the fatty acid is not split away from them to any major extent Consequently, the purpose of their being in the cells of the brain is presumably related to their indestructible physical properties.

ATHEROSCLEROSIS

Atherosclerosis is a disease of the intima of the arteries, especially of the large arteries, that leads to fatty lesions called *atheromatous plaques* on the inner surfaces of the arteries. The earliest stage in the development of these lesions is believed to be damage to the endothelial cells and sublying intima. The damage can be caused by physical abrasion of the endothelium, by abnormal substances in the blood, or even by the effect of the pulsating arterial pressure on the vessel wall Once the damage has occurred, the endothelial cells swell and proliferate, and even sublying smooth muscle cells proliferate and migrate from the media of the arteries into the lesion Soon thereafter lipid substances, especially cholesterol, begin to deposit from the blood in the proliferating cells, forming the atheromatous plaques Because these plaques contain so much cholesterol, they are frequently called simply *cholesterol deposits* In the later stages of the lesions, fibroblasts infiltrate the degenerative areas and cause progressive sclerosis (fibrosis) of the arteries And still later, calcium often precipitates with the lipids to develop *calcific plaques* When these two processes have occurred, the arteries are then extremely hard, and the disease is called *arteriosclerosis*, or simply "hardening of the arteries"

Obviously, arteriosclerotic arteries lose most of their distensibility, and because of the degenerative areas they are easily ruptured Also, the atheromatous plaques often protrude into the flowing blood, and the roughness of their surfaces causes blood clots to develop, with resultant thrombus or embolus formation (see Chapter 9) Almost half of all human beings in the United States and Europe die of arteriosclerosis. Approximately two thirds of these deaths are caused by thrombosis of one or more coronary arteries, and the remaining one third by thrombosis or hemorrhage of vessels in other organs of the body—especially the brain, that causes strokes—but also in the kidneys, liver, gastrointestinal tract, limbs, and so forth

Despite the extreme prevalence of atherosclerosis, little is known about its cause Therefore, it is necessary to outline the general trends of the experimental studies rather than to present a definitive description of the mechanisms that cause atherosclerosis.

EXPERIMENTAL PRODUCTION OF ATHEROSCLEROSIS IN ANIMALS

In the past, it had been believed that all that was necessary to cause atherosclerosis was to increase the amount of cholesterol circulating in the blood Indeed, very early it was demonstrated that feeding large amounts of cholesterol to rabbits will cause large numbers of cholesterol deposits in the intimal layer of the arteries This experiment has often been cited as proof of the hypercholesterolemic theory of atherogenesis However, these deposits do not lead to subsequent arterial wall fibrosis and death of the animal Furthermore, it has been difficult to achieve such deposits in carnivorous animals such as the dog, except by feeding extreme amounts of cholesterol and also removing the thyroid gland to prevent normal utilization of cholesterol.

Therefore, in recent years much more emphasis has been placed on the initial endothelial and intimal lesions as the primary cause of the ultimate atherosclerotic plaques. Almost any factor that can cause damage to the endothelial cells will lead to the following sequence of events First, platelets adhere to the endothelium Second, the platelets dissolute, and some factor from the platelets causes proliferation of the sublying smooth muscle cells; these then infiltrate the damaged region

Subsequently, cholesterol (and to much lesser exteot other lipids) infiltrates the lesion until eveotually lesions typical of human atherosclerotic plaques develop. The severity of these lesions is enhanced in the presence of hypercholesterolemia.

ATHEROSCLEROSIS IN THE HUMAN BEING

Effect of Age, Sex, and Heredity on Atherosclerosis. Atherosclerosis is mainly a disease of old age, but small atheromatous plaques can almost always be found in the arteries of young adults. Therefore, the full-blown disease is a culminatioo of a lifetime of vascular damage and lipid deposition rather than deposition over a few years.

Far more men than women die of atherosclerotic heart disease. This is especially true of men younger than 50 years of age. For this reason, it is possible that the male sex hormone accelerates the development of atherosclerosis, or that the female sex hormone *protects* a person from atherosclerosis. Indeed, admioistration of estrogens to men who have already had coronary thromboses has decreased the number of secondary coronary attacks in some clinical trials. Furthermore, administration of estrogens to chickens with atheromatous plaques in their coronaries has actually caused the disease to regress.

Atherosclerosis and atherosclerotic heart disease are highly hereditary in some families. In some instances, this is related to an inherited *familial hypercholesterolemia*, the excess cholesterol occurring almost entirely in the *low density lipoproteins*. This most often results from lack of *low density lipoprotein receptors* on the cell membrenes throughout the body. These receptors recognize a protein called *apoprotein B-100* of the low density llpoproteins and causes the lipoproteins to adhere to the cells. This adherence is required before the lipoproteins can deliver their load of cholesterol to the cells, especially back to the liver cells from which the cholesterol had come in the first place In the absence of the return of most of the cholesterol to the liver cells, the internal feedback mechanism of the llver cells causes prolific production of cholesterol along with the formation of still additional low density lipoproteins. Therefore, the blood conceotration of cholesterol, found principally in low density lipoproteins, can rise to as high as 600 mg/dl, more than three times the normal level of 175 to 200 mg/dl. Homozygous persons with this disease usually die of a heart attack before the age of 20 years.

In other persons with hereditary atherosclerosis, the blood cholesterol level is completely normal, but other, less well-defined factors lead to atherosclerosis nevertheless.

Other Factors That Predispose to Atherosclerosis. Human beings with severe *diabetes* or severe *hypothyroidism* frequently develop premature and severe atherosclerosis. In both these conditions the blood cholesterol is greatly elevated, which is at least part of the cause of the atherosclerosis.

Another disease associated with atherosclerosis, in human beings as well as in experimental animals, is *hypertension;* the incidence of atherosclerotic coronary heart disease is about twice as great in hypertensive people as in normal persons. Though the cause is not known, it possibly results from pressure damage to the arterial walls, with subsequent deposition of cholesterol plaques.

Still another factor that increases the likelihood of atherosclerosis is smoking, though here again the cause of this relationship is not understood. On the average, a person who smokes approximately doubles his or her chances of dying from a heart attack.

Relationship of Dietary Fot to Atherosclerosis in the Human Being. A high fat diet, especially one containing cholesterol and saturated fats, increases one's chances of developing atherosclerosis. Therefore, decreasing the fat can help greatly in protecting against atherosclerosis, and some experiments indicate that this can benefit even patients who have already had coronary heart attacks A recent clinical trial conducted by the United States National Institutes of Health on dietary measures to reduce the blood cholesterol level showed that for each 1 per cent decrease in the cholestcrol there was about a 2 per cent decrease in mortality from heart attacks Also, life insurance statistics show that the rate of mortality—mainly from coronary disease—of normal weight middle and older aged persons is about half the mortality rate of overweight subjects of the same age.

As pointed out earlier, eating unsaturated fats instead of saturated fats, for reasons not known, can decrease the blood cholesterol concentration, and this seems to be beneficial in preventing atherosclerosis

SUMMARY OF FACTORS CAUSING ATHEROSCLEROSIS

Atherosclerosis is highly associated with abnormalities of lipid metabolism, but it is also exacerbated by almost any factor that injures the arterial wall In particular, elevated blood cholesterol is often related to atherosclerosis. But, perhaps equally as important might be some undiscovered third factor that is inherited from generation to generation which causes increased arterial intimal degeneration or increased rate of cholesterol deposition in the arterial walls irrespective of the blood cholesterol concentration

REFERENCES

Bendit, E P The origin of atherosclerosis *Sci Am*, 236(2) 74, 1977

Bisgaier, C, and Glickman, R M Intestinal synthesis, secretion, and transport of lipoproteins *Annu Rev Physiol*, *45* 625, 1983

Brady, G A, and York, D A Hypothalame and genetic obesity in experimental animals An autonomic and endocrine hypothesis *Physiol Rev*, *59* 719, 1979

Brady, R O The sphingolipidoses *In* Bondy, P K and Rosenberg, L E (eds.) Metabolic Control and Disease, 8th Ed *Philadelphia*, *W B* Saunders Co, 1980, p 523

Bremer, J Carnitine—metabolism and functions *Physiol Rev*, *63* 1420, 1983

Brownell, K D The psychology and physiology of obesity Implications for screening and treatment *J Am Diet Assoc*, *84* 406, 1984

Bukowiecki, L J Mechanisms of stimulus-calorigenesis coupling in brown adipose tissue *Can J Biochem Cell Biol* *62* 623, 1984

Carey, M C, *et al* Lipid digestion and absorption. *Annu Rev Physiol*, *45* 651, 1983

Chen, L. Hormonal control of apolipoprotein synthesis. *Annu Rev Physiol*, *45* 615, 1983

Chandler, A B, *et al* (eds) The Thrombotic Process in Atherogenesis New York, Plenum Press, 1978

Coleman, J E: Metabolic interrelationships between carbohydrates, lipids and proteins *In* Bondy, P K., and Rosenberg, L E (eds): Metabolic Control and Disease, 6th Ed Philadelphia, W B Saunders, Co, 1980, p 161

Crepaldi, G, *et al* (eds.). Diabetes, Obesity, and Hyperlipidemias. New York, Academic Press, 1978

Dockerman, R. M. Gastric exclusion surgery in the management of morbid obesity. *Annu Rev Med.*, *33* 263, 1982

Dole, R., and Knudsen, J (eds) Regulation of Fatty Acid and Glycerolipid Metabolism New York, Pergamon Press, 1978

Fitzgerald F T The problem of obesity *Annu Rev Physiol*, *32* 21, 1981
Forte, T M Primary hepatocytes in monolayer culture A model for studies on lipoprotein metabolism *Annu Rev Physiol*, *46* 403 1984
Fumagalli (ed) Drugs Affecting Lipid Metabolism New York, Elsevier Science Publishing Co, Inc, 1981
Gennis, R B, and Jonas A Protein Lipid Interactions *Annu Rev Biophys Bioeng* 6 195, 1977
Goldfarb, S Regulation of hepatic cholesterogenesis *In* Javitt, N B (ed) International Review of Physiology Liver and Biliary Tract Physiology I Vol 21 Baltimore, University Park Press, 1980, p 317
Golinick, P D Metabolism of substrates. Energy substrate metabolism during exercise and as modified by training *Fed Proc*, *44* 353, 1985
Goto, Y New aspects of lipid metabolism and arteriosclerosis *In* Hayase, S, and Murao, S (eds) Cardiology Proceedings of the VIII World Congress of Cardiology, Tokyo, 1978 New York, Elsevier/North Holland, 1979, p 197
Gotto, A M, Jr, *et al* (eds) High Density Lipoproteins and Atherosclerosis New York, Elsevier/North Holland, 1978
Gresham, G A Reversing Atherosclerosis Springfield, Ill, Charles C Thomas, 1979
Grundy, S M Absorption and metabolism of dietary cholesterol Annu *Rev Nutr*, *3* 71, 1983
Hashimoto S, and Dayton, S Lipid metabolism of arterial smooth muscle *In* Bohr, D F, *et al* (eds) Handbook of Physiology Sec 2 Vol II Baltimore Williams & Wilkins 1980 p 161
Havel, R J, and Kane, J P Therapy of hyperlipidemic states *Annu Rev Med* 33 417, 1982
Havel, R J *et al* Lipoproteins and lipid transport. *In* Bondy P K, and Rosenberg L E (eds) Metabolic Control and Disease 8th Ed Philadelphia, W B Saunders Company, 1980 p 393
Hessel L W and Kraus, H M J (eds) Lipoprotein Metabolism and Endocrine Regulation New York, Elsevier/North Holland 1979
Hietanan, E Regulation of Serum Lipids by Physical Exercise Boca Raton, FL, CRC Press, 1982
Horrocks, L A (ed) Phospholipids in the Nervous System Metabolism New York, Raven Press 1982
Jackson R L, *et al* Lipoprotein structure and metabolism *Physiol Rev*, *56* 259 1976
Kane, J P Apolipoprotein B Structural and metabolic heterogeneity *Annu Rev Physiol*, *45* 637, 1983
Levy, R I (eds) Nutrition, Lipids, and Coronary Heart Disease New York Raven Press 1979
Masoro, E J Lipids and lipid metabolism *Annu Rev Physiol*, *39* 301, 1977
McGill H. C, Jr Morphologic development of the atherosclerotic plaque *In* Lauer, R M and Shekele R B (eds) Childhood Prevention of Atherosclerosis and Hypertension New York Raven Press 1980
Miller G J High density lipoproteins and atherosclerosis *Annu Rev Med* *31* 97 1980
Nelson, G J (ed) Blood Lipids and Lipoproteins Quantitation, Composition and Metabolism Huntington, N Y R E Krieger Publishing Co, 1979
Parmley, W W, *et al*: Modification of experimental atherosclerosis calcium-channel blockers *Am J Cardiol.*, *55* 165B, 1985
Peng, S K, and Taylor, C B., Cholesterol autoxidation, health and arteriosclerosis A review on situations in developed countries. *World Rev Nutr Diet*, *44* 117, 1984
Robertson, A L, Jr, and Rosen, L A The endothelial lining of arteries in disease *In* Kaley, G, and Altura, B M (eds) Microcirculation Vol III Baltimore University Park Press, 1977
Robinson, A M, and Williamson, D H Physiological roles of ketone bodies as substrates and signals in mammalian tissues *Physiol Rev* *60* 143, 1980
Rosell, S Microcirculation and transport in adipose tissue *In* Renkin E M, and Michel, C C. (eds) Handbook of Physiology Sec. 2 Vol IV Bethesda American Physiological Society, 1984, p 949
Rosell, S. and Belfrage, E Blood circulation in adipose tissue *Physiol Rev*, *59* 1078, 1979
Rosenberg R D Role of heparin and heparinlike molecules in thrombosis and atherosclerosis. *Fed Proc*, *44* 404, 1985
Ross, R Endothelial integrity, cell proliferation, and atherosclerosis Lauer, R M, and Shekelle, R B (eds) Childhood Prevention of Atherosclerosis and Hypertension New York, Raven Press, 1980
Ross R., and Kariya, B Morphogenesis of vascular smooth muscle in atherosclerosis and cell structure *In* Bohr, D R, *et al* (eds) Handbook of Physiology Sec 2, Vol II Baltimore, Williams & Wilkins 1980, p 69
Royce, S M, *et al* The influence of dietary isomeric and saturated fatty acids on atherosclerosis and eicosanoid synthesis in swine *Am J Clin Nutr*, *39* 215, 1984
Salans L B Obesity and the adipose cell *In* Bondy, P. K, and Rosenberg, L. E (eds) Metabolic Control and Disease, 8th Ed Philadelphia W B Saunders Co, 1980 p 495
Samuel, P, *et al* The role of diet in the etiology and treatment of atherosclerosis *Annu Rev Med*, *34* 179, 1983
Sesnu, A M *et al* (eds) The Biochemistry of Atherosclerosis New York, Marcel Dekker, 1978
Schonfeld, G Hormonal control of lipoprotein metabolism *In* DeGroot L J *et al* (eds) Endocrinology Vol 3 New York, Grune & Stratton 1979, p 1855
Stallones, R A Ischemic heart disease and lipids in blood and diet *Annu Rev Nutr* *3* 155, 1983
Szoka F, Jr, and Papahadjopoulos, D Comparative properties and methods of preparation of lipid vesicles (liposomes), Annu Rev Biophys Bioeng, *9* 467, 1980
Texon, M The Hemodynamic Basis of Atherosclerosis Washington, D C, Hemisphere Publishing Corp, 1979
White, F N, and Somero, G Acid base regulation and phospholipid adaptations to temperature Time courses and physiological significance of modifying the milieu for protein function *Physiol Rev*, 62 40 1982
Wight, T N Proteoglycans in pathological conditions Atherosclerosis *Fed Proc*, *44* 381, 1985
Wissler, R W Nutrition plasma lipids, and atherosclerosis. *In* Lauer, R M, and Shekelle, R B (eds) Childhood Prevention of Atherosclerosis and Hypertension New York, Raven Press, 1980

69

Protein Metabolism

About three quarters of the body solids are proteins. These include *structural proteins, enzymes, nucleoproteins, proteins that transport oxygen, proteins of the muscle that cause contraction,* and many other types that perform specific functions both intracellularly and extracellularly throughout the body

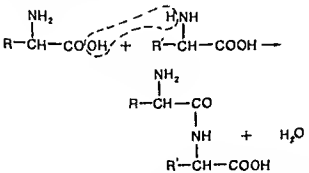
The basic chemical properties of proteins that explain their diverse functions are so extensive that they constitute a major portion of the entire discipline of biochemistry. For this reason, the present discussion is confined to a few specific aspects of protein metabolism that are important as background for other discussions in this text.

BASIC PROPERTIES OF PROTEINS

THE AMINO ACIDS

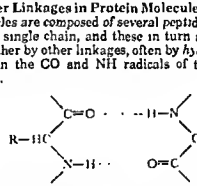
The principal constituents of proteins are amino acids, 20 of which are present in the body proteins in significant quantities Figure 69–1 illustrates the chemical formulas of these 20 amino acids, showing that they all have two features in common. Each amino acid has an acidic group (—COOH) and a nitrogen radical attached to the molecule near the acidic radical, usually represented by the amino group ($—NH_2$)

Peptide Linkages and Peptide Chains. In proteins, the amino acids are aggregated into long chains by means of *peptide linkages,* one of which is illustrated by the following reaction:



Note in this reaction that the nitrogen of the amino radical of one amino acid bonds with the carbon of the carboxyl radical of the other amino acid. A hydrogen atom is released from the amino radical, and a hydroxyl radical is released from the carboxyl radical; these two combine to form a molecule of water. Note that after the peptide linkage has been formed, an amino radical and a carboxyl radical are still in the new molecule, both of which are capable of combining with additional amino acids to form a *peptide chain* Some complicated protein molecules have many thousand amino acids combined together principally by peptide linkages, and even the smallest protein usually has more than 20 amino acids combined together by peptide linkages The average is about 400 amino acids

Other Linkages in Protein Molecules. Some protein molecules are composed of several peptide chains rather than a single chain, and these in turn are bound with each other by other linkages, often by *hydrogen bonding* between the CO and NH radicals of the peptides as follows.



Also, many peptide chains are coiled or folded, and the successive coils or folds are held in a tight spiral or in other shapes by similar hydrogen bonding But, in addition to hydrogen bonds, separate peptide chains can be held together by hydrophobic bonds, electrostatic forces, and sulfhydryl, phenolic, and salt linkages, as well as by others

TRANSPORT AND STORAGE OF AMINO ACIDS

THE BLOOD AMINO ACIDS

The normal concentration of amino acids in the blood is between 35 and 65 mg/dl. This is an average of about 2 mg/dl for each of the 20 amino acids, though some are present in far greater concentrations than others Since the amino acids are relatively strong acids, they exist in the blood principally in the ionized state and account for 2 to 3 milliequivalents of the negative ions in the blood. The precise distribution of the different amino acids in the blood depends to some extent on the types of proteins ingested, but the concentrations of at least some individual amino acids are regulated by selective synthesis in the different cells.

Fate of Amino Acids Absorbed from the Gastrointestinal Tract. It will be recalled from Chapter 65 that

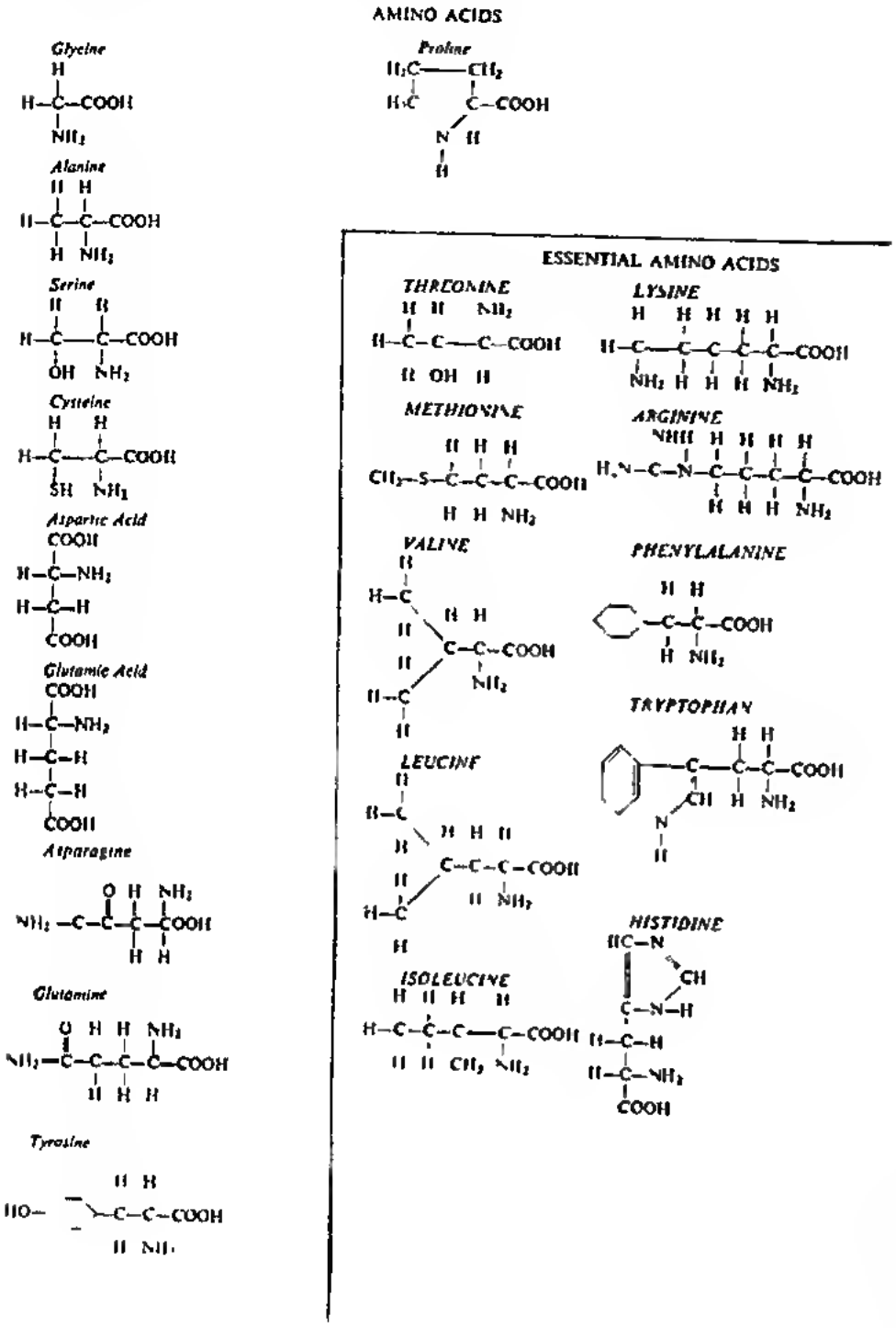


Figure 69–1. The amino acids, showing the 10 essential amino acids, which cannot be synthesized at all or in sufficient quantity in the body

the end-products of protein digestion in the gastrointestinal tract are almost entirely amino acids and that only rare polypeptide or whole protein molecules are absorbed. Immediately after a meal, the amino acid concentration in the blood rises, but the rise is usually only a few milligrams per cent for two reasons: First, protein digestion and absorption is usually extended over two to three hours, which allows only small quantities of amino acids to be absorbed at a time. Second, after entering the blood, the excess amino acids are absorbed within five to ten minutes by cells throughout the entire body, especially by the liver. Therefore, almost never do large concentrations of amino acids accumulate in the blood. Nevertheless, the turnover rate of the amino acids is so rapid that many grams of proteins can be carried from one part of the body to another in the form of amino acids each hour.

Active Transport of Amino Acids into the Cells. The molecules of essentially all the amino acids are much too large to diffuse through the pores of the cell membranes. Therefore, significant quantities of amino acids can be transported through the membrane only by

facilitated or active transport utilizing carrier mechanisms. The nature of the carrier mechanisms is still poorly understood, but some are discussed in Chapter 9.

Renal Threshold for Amino Acids. One of the special functions of carrier transport of amino acids is to prevent loss of these in the urine. All the different amino acids can be *actively transported* through the proximal tubular epithelium, thus removing them from the glomerular filtrate and returning them to the blood However, as is true of other active transport mechanisms in the renal tubules, there is an upper limit to the rate at which each type of amino acid can be transported. For this reason, when a particular type of amino acid rises to too high a concentration in the plasma and glomerular filtrate, the excess above that which can be *actively reabsorbed* is lost into the urine

In Chapter 38 it was pointed out that appropriate carrier systems for active reabsorption of certain amino acids are often deficient or lacking in the renal tubular epithelium. Under these conditions the plasma threshold for the respective amino acids is greatly reduced However, in the normal person, the loss of amino acids in the urine each day is insignificant

STORAGE OF AMINO ACIDS AS PROTEINS IN THE CELLS

Almost immediately after entry into the cells, amino acids are combined by peptide linkages, under the direction of the messenger RNA and ribosomal system, to form cellular proteins so that the concentrations of the amino acids inside the cells probably always remains low. Thus, storage of large quantities of amino acids as such probably does not occur in the cells; instead, they are mainly stored in the form of actual proteins. Yet many intracellular proteins can be rapidly decomposed again into amino acids under the influence of intracellular lysosomal digestive enzymes, and these amino acids in turn can be transported back out of the cell into the blood. The proteins that can thus be decomposed include many cellular enzymes as well as some other functioning proteins. However, the proteins in the chromosomes of the nucleus and the structural proteins such as collagen and muscle contractile proteins do not participate significantly in this reversible storage of amino acids

Some tissues of the body participate in the storage of amino acids to a greater extent than others. For instance, the liver, which is a large organ and also has special systems for processing amino acids, can store large quantities of proteins; this is also true of the kidney and the intestinal mucosa.

Release of Amino Acids from the Cells and Regulation of Plasma Amino Acid Concentration. Whenever the plasma amino acid concentrations fall below their normal levels, amino acids are transported out of the cells to replenish the supply in the plasma. Simultaneously, intracellular proteins are degraded back into amino acids.

The plasma concentration of each type of amino acid is maintained at a reasonably constant value. Later it will be noted that the various hormones secreted by the endocrine glands are able to alter the balance between tissue proteins and circulating amino acids; growth hormone and insulin increase the formation of tissue proteins, while the adrenocortical glucocorticoid hormones increase the concentration of circulating amino acids.

Reversible Equilibrium Between the Proteins of Different Parts of the Body. Since cellular proteins in the liver (and to a much lesser extent in other tissues) can be synthesized rapidly from plasma amino acids and many of these in turn can be degraded and returned to the plasma almost equally as rapidly, there is constant equilibrium between the plasma amino acids and most of the proteins in the cells of the body. Therefore, it follows that there is also equilibrium between the proteins from one type of cell and the next. For instance, if any particular tissue loses proteins, it can synthesize new proteins from the amino acids of the blood; in turn, these are replenished by degradation of proteins from other cells of the body. These effects are particularly noticeable in relation to protein synthesis in cancer cells Cancer cells are prolific users of amino acids, and, simultaneously, the proteins of the other tissues become markedly depleted.

Upper Limit to the Storage of Proteins. Each particular type of cell has an upper limit to the amount of proteins that it can store. After all the cells have reached their limits, the excess amino acids in the circulation are then degraded into other products and used for energy, as is discussed subsequently, or they are converted to fat or glycogen and stored in these forms.

METABOLIC ROLES OF THE PLASMA PROTEINS

The three major types of protein present in the plasma are *albumin, globulin,* and *fibrinogen* The principal function of albumin is to provide *colloid osmotic pressure,* which in turn prevents plasma loss from the capillaries, as discussed in Chapter 30 The globulins perform a number of enzymatic functions in the plasma itself, but, more important than this, they are principally responsible for both the natural and acquired immunity that a person has against invading organisms, which was discussed in Chapter 6 The fibrinogen polymerizes into long fibrin threads during blood coagulation, thereby forming blood clots that help to repair leaks in the circulatory system, which was discussed in Chapter 8

Formation of the Plasma Proteins. Essentially all the albumin and fibrinogen of the plasma proteins, as well as 60 to 80 per cent of the globulins, are formed in the liver. The remainder of the globulins are formed in the lymphoid tissues. These are mainly the gamma globulins that constitute the antibodies

The rate of plasma protein formation by the liver can be extremely high, as much as 50 grams per day. Certain disease conditions often cause rapid loss of plasma proteins; severe burns that denude large surface areas of the skin cause loss of many liters of plasma through the denuded areas each day. The rapid production of plasma proteins by the liver is obviously valuable in preventing death in such states. Furthermore, occasionally, a person with severe renal disease loses as much as 20 grams of plasma protein in the urine each day for months, and these are continually replaced.

Use of Plasma Proteins by the Tissues. When the tissues become depleted of proteins, the plasma proteins can act as a source for rapid replacement of the tissue proteins. Indeed whole plasma proteins [illegible] *in toto* by the tissue macrophages by [illegible]

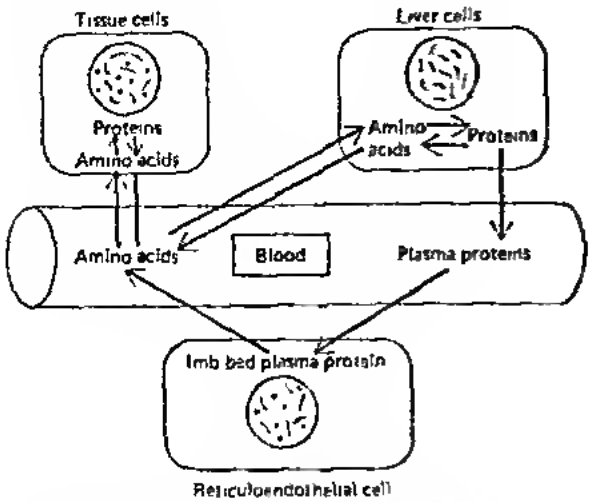


Figure 69–2. Reversible equilibrium among the tissue proteins, plasma proteins and plasma amino acids

pinocytosis, then, once in the cells, these are split into amino acids that are transported back into the blood and utilized throughout the body to build cellular proteins. In this way, then, the plasma proteins function as a labile protein storage medium and represent a rapidly available source of amino acids whenever a particular tissue requires these.

Reversible Equilibrium Between the Plasma Proteins and the Tissue Proteins. The rate of synthesis of plasma proteins by the liver depends on the concentration of amino acids in the blood, which means that the concentration of plasma proteins becomes reduced whenever an appropriate supply of amino acids is not available. On the other hand, whenever excess proteins are available in the plasma but insufficient proteins are present in the cells, the plasma proteins are used to form tissue proteins. Thus, there is a constant state of equilibrium, as illustrated in Figure 69–2, between the plasma proteins, the amino acids of the blood, and the tissue proteins. It has been estimated from radioactive tracer studies that about 400 grams of body protein are synthesized and degraded each day as part of the continual state of flux of amino acids. This illustrates once again the general principle of reversible exchange of amino acids among the different proteins of the body. Even during starvation or during severe debilitating diseases, the ratio of total tissue proteins to total plasma proteins in the body remains relatively constant at about 33 to 1.

Because of this reversible equilibrium between plasma proteins and the other proteins of the body, one of the most effective of all therapies for severe acute protein deficiency is intravenous administration of plasma protein. Within a few days or sometimes within hours, the amino acids of the administered protein become distributed throughout the cells of the body to form proteins where they are needed.

ESSENTIAL AND NONESSENTIAL AMINO ACIDS

Ten of the amino acids normally present in animal proteins can be synthesized in the cells, whereas the other ten either cannot be synthesized at all or are synthesized in quantities too small to supply the body's needs. The first group of amino acids is called *nonessential*, while the second group is called *essential amino acids*. The essential amino acids obviously must be present in the diet if protein formation is to take place in the body. Use of the word "essential" does not mean that the other ten amino acids are not equally essential for formation of the proteins, but only that these others are not essential in the diet.

Synthesis of the nonessential amino acids depends on the formation first of appropriate α-keto acids, which are the precursors of the respective amino acids. For instance, *pyruvic acid*, which is formed in large quantities during the glycolytic breakdown of glucose, is the keto acid precursor of the amino acid *alanine*. Then, by the simple process of *transamination*, an amino radical is transferred to the α-keto acid while the keto oxygen is transferred to the donor of the amino radical. The reaction is illustrated in Figure 69–3. Note in the reaction that the amino radical is transferred to the pyruvic acid from another amino acid, *glutamine*. Glutamine is present in the tissues in large quantities, and it functions specifically as an amino radical storehouse. In addition, amino radicals can be transferred from *asparagine*, *glutamic acid*, and *aspartic acid*.

Transamination is promoted by enzymes called *transaminases*, all of which are derivatives of pyridoxine, one of the B vitamins. Without this vitamin, the nonessential amino acids are not synthesized at all or only poorly so, and, therefore, protein formation cannot proceed normally.

USE OF PROTEINS FOR ENERGY

It was pointed out earlier that there is an upper limit to the amount of protein that can accumulate in each particular type of cell. Once the cells are filled to the limits, any additional amino acids in the body fluids are degraded and used for energy or stored as fat. The degradation occurs almost entirely in the liver, and it begins with the process known as *deamination*. To help in this process, the excess amino acids in the cells, especially in the liver, induce the production of large quantities of *aminotransferases*, the enzymes responsible for initiating most deamination.

Deamination. Deamination means removal of the amino groups from the amino acids. This can occur

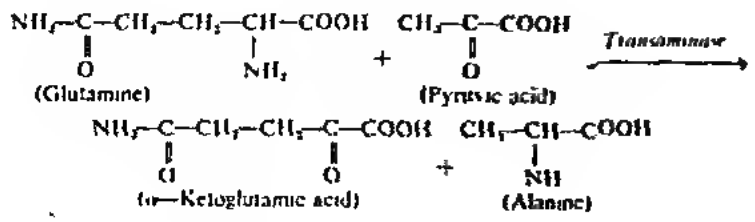
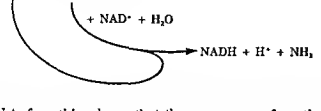


Figure 69–3. Synthesis of alanine from pyruvic acid by transamination.

several different means, two of which are especially important: (1) transamination, which means transfer of the amino group to some acceptor substance as explained above in relation to the synthesis of amino acids, and (2) oxidative deamination.

The greatest amount of deamination occurs by the following transamination schema:



Note from this schema that the amino group from the amino acid is transferred to α-ketoglutaric acid, which then becomes glutamic acid. The glutamic acid can then transfer the amino group to still other substances or can release it in the form of ammonia. In the process of losing the amino group, the glutamic acid once again becomes α-ketoglutaric acid, so that the cycle can be repeated again and again.

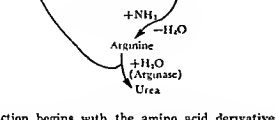
Oxidative deamination occurs to much less extent and is catalyzed by amino acid oxidases. In this process the amino acid is oxidized at the point where the amino radical attaches, which causes the amino radical to be released.

Urea Formation by the Liver. The ammonia released during deamination is removed from the blood almost entirely by conversion into urea, two molecules of ammonia and one molecule of carbon dioxide combining in accordance with the following net reaction:

$$2NH_3 + CO_2 \rightarrow H_2N{-}\underset{\underset{O}{\|}}{C}{-}NH_2 + H_2O$$

Essentially all urea formed in the human body is synthesized in the liver. In the absence of the liver or in serious liver disease, ammonia accumulates in the blood. This in turn is extremely toxic, especially to the brain, often leading to a state called *hepatic coma.*

The stages in the formation of urea are essentially the following:



The reaction begins with the amino acid derivative *ornithine,* which combines (by means of a series of reactions) with one molecule of carbon dioxide and one molecule of ammonia to form a second substance, *citrulline.* This in turn combines with still another molecule of ammonia to form *arginine,* which then splits into *ornithine* and *urea.* The urea diffuses from the liver cells into the body fluids and is excreted by the kidneys, while the ornithine is reused in the cycle again and again.

Oxidation of Deaminated Amino Acids. Once amino acids have been deaminated, the resulting keto acid products can in most instances be oxidized to release energy for metabolic purposes. This usually involves two successive processes. (1) the keto acid is changed into an appropriate chemical substance that can enter the citric acid cycle, and (2) this substance is then degraded by this cycle and used for energy in the same manner that acetyl-CoA derived from carbohydrate and lipid metabolism is used.

In general, the amount of adenosine triphosphate formed for each gram of protein that is oxidized is slightly less than that formed for each gram of glucose oxidized.

Gluconeogenesis and Ketogenesis. Certain deaminated amino acids are similar to the breakdown products that result from glucose and fatty acid metabolism. For instance, deaminated alanine is pyruvic acid. Obviously, this can be converted into glucose or glycogen. Or it can be converted into acetyl-CoA, which can then be polymerized into fatty acids. Also, two molecules of acetyl-CoA can condense to form acetoacetic acid, which is one of the ketone bodies, as explained in the preceding chapter.

The conversion of amino acids into glucose or glycogen is called *gluconeogenesis,* and the conversion of amino acids into keto acids or fatty acids is called *ketogenesis.* Eighteen of 20 of the deaminated amino acids have chemical structures that allow them to be converted into glucose, and 19 can be converted into fats—5 directly and the other 14 by becoming carbohydrate first and then becoming fat.

OBLIGATORY DEGRADATION OF PROTEINS

When a person eats no proteins, a certain proportion of that person's own body proteins continues to be degraded into amino acids, then deaminated and oxidized. This involves 20 to 30 grams of protein each day, which is called the *obligatory loss* of proteins. Therefore, to prevent a net loss of protein from the body, one must ingest a minimum of 20 to 30 grams of protein each day, and to be on the safe side a minimum of 60 to 75 grams is usually recommended.

The ratios of the different amino acids in the dietary protein must be about the same as the ratios in the body tissues if the entire protein is to be usable. If one particular type of essential amino acid is low in concentration, the others become unusable as well because cells form either whole proteins or none at all, as was explained in Chapter 3 in relation to protein synthesis. The unusable amino acids are then deaminated and oxidized. A protein that has a ratio of amino acids *different from* that of the *average body protein is called* a *partial protein* or *incomplete protein,* and such a protein is obviously less valuable for nutrition than is the *complete protein.*

Effect of Starvation on Protein Degradation. Except for the excess protein in the diet or the 20 to 30 grams of obligatory protein degradation each day, the body uses almost entirely carbohydrates or fats for energy as long as these are available. However, after

several weeks of starvation, when the quantity of stored fats begins to run out, the amino acids of the blood begin to be rapidly deaminated and oxidized for energy. From this point on, the proteins of the tissues degrade rapidly—as much as 125 grams daily—and the cellular functions deteriorate precipitously

Because carbohydrate and fat utilization for energy occurs in preference to protein utilization, carbohydrates and fats are called *protein sparers*

HORMONAL REGULATION OF PROTEIN METABOLISM

Growth Hormone. Growth hormone increases the rate of synthesis of cellular proteins, causing the tissue proteins to increase The precise mechanism by which growth hormone increases the rate of protein synthesis is not known, but growth hormone is believed to act in some direct manner to enhance the transport of amino acids through the cell membranes and/or to accelerate the DNA and RNA processes of protein synthesis Part of the action might also result from the effect of growth hormone on fat metabolism, for this hormone causes increased rate of fat liberation from the fat depots, making this available for energy. This in turn reduces the rate of oxidation of amino acids and consequently makes increased quantities of amino acids available to the tissues to be synthesized into proteins

Insulin. Lack of insulin reduces protein synthesis almost to zero The mechanism by which this hormone affects protein metabolism is also unknown, but insulin does accelerate amino acid transport into the cells, which could be the stimulus to protein synthesis Also, insulin increases the availability of glucose to the cells so that the use of amino acids for energy becomes correspondingly reduced This, too, undoubtedly makes far larger quantities of amino acids available to the tissues for protein synthesis

Glucocorticoids. The glucocorticoids secreted by the adrenal cortex *decrease* the quantity of protein in most tissues while increasing the amino acid concentration in the plasma However, contrary to elsewhere in the body, these hormones increase both the liver proteins and the plasma proteins It is believed that the glucocorticoids act by increasing the rate of breakdown of extrahepatic proteins, thereby making increased quantities of amino acids available in the body fluids This in turn supposedly allows the liver to synthesize increased quantities of hepatic cellular proteins and plasma proteins.

The effects of glucocorticoids on protein metabolism are especially important in ketogenesis and gluconeogenesis, for in the absence of these hormones insufficient quantities of amino acids are usually available in the plasma to allow either significant gluconeogenesis or ketogenesis from proteins

Testosterone. Testosterone, the male sex hormone, causes increased deposition of protein in the tissues throughout the body, including especially increase in the contractile proteins of the muscles. The mechanism by which this effect comes about is unknown, but it is definitely different from the effect of growth hormone in the following way: Growth hormone causes tissues to continue growing almost indefinitely, while testosterone causes the muscles and other protein tissues to enlarge ly for several months, beyond that time, despite con- administration of testosterone, further protein ceases.

Estrogen, the principal female sex hormone, also causes slight deposition of protein, but its effect is relatively insignificant in comparison with that of testosterone.

Thyroxine. Thyroxine increases the rate of metabolism of all cells, and, as a result, indirectly affects protein metabolism If insufficient carbohydrates and fats are available for energy, thyroxine causes rapid degradation of proteins and uses these for energy. On the other hand, if adequate quantities of carbohydrates and fats are available and excesses of amino acids are also available in the extracellular fluid, thyroxine can actually increase the rate of protein synthesis. Conversely, in growing animals deficiency of thyroxine causes growth to be greatly inhibited because of lack of protein synthesis In essence, it is believed that thyroxine has little specific direct effect on protein metabolism but does have an important general effect in increasing the rates of both normal anabolic and normal catabolic protein reactions

REFERENCES

Atkinson D E, and Fox C F (eds) Modulation of Protein Function. New York, Academic Press, 1979

Benevenga, N J, and Steele, R D Adverse effects of excessive consumption of amino acids *Annu. Rev Nutr*, *4* 157, 1984

Borum, P R. Carnitine *Annu. Rev Nutr*, *3* 233, 1983

Bremer, H J *et al* Clinical Chemistry and Diagnosis of Amino Acid Disturbances Baltimore, Urban & Schwarzenberg, 1979

Christensen, H N Interorgan amino acid nutrition. *Physiol. Rev.*, *62* 1193, 1982

Coleman, J E Metabolic interrelationships between carbohydrates, lipids and proteins *In* Bondy, P K and Rosenberg, L. E. (eds.) Metabolic Control and Disease, 8th Ed Philadelphia, W B Saunders Co 1980, p 161

Dohm, G L. *et al* Protein metabolism during endurance exercise *Fed. Proc*, *44* 348, 1985

Foster, R L. The Nature of Enzymology New York, John Wiley & Sons, 1979

Franks, F (ed) Characterization of Protein Conformation and Function. London, Symposium Press, 1978

Friedmann, H C (ed) Enzymes. Stroudsburg Pa, Dowden, Hutchinson & Ross, 1980

Galjaard, H Genetic Metabolic Diseases Early Diagnosis and Prenatal Analysis New York, Elsevier/North-Holland, 1980

Gevers W Protein metabolism of the heart. *J Mol Cell Cardiol*, *16* 3, 1984

Gross, E, and Meienhofer, J (eds) The Peptides. New York, Academic Press, 1979

Harper A E *et al* Branched chain amino acid metabolism *Annu. Rev Nutr*, *4* 409, 1984

Kinney J M, and Elwyn D H Protein metabolism and injury *Annu Rev Nutr*, *3* 433, 1983

Ormsdan, J L Recent Advances in Endocrinology and Metabolism. New York, Churchill Livingstone, 1982

Rosenberg L. E, and Scriver, C R. Disorders of amino acid metabolism *In* Bondy, P K, and Rosenberg, L E (eds.) Metabolic Control and Disease, 8th Ed. Philadelphia, W B Saunders Co, 1980, p 583

Rothschild, M A Albumin synthesis. *In* Javitt, N B (ed) International Review of Physiology Liver and Biliary Tract Physiology I. Vol 21 Baltimore, University Park Press, 1980, p 249

Rothschild, M A, *et al* Effect of albumin concentration on albumin synthesis in the perfused liver *Am J Physiol*, *216* 1127, 1969

Seegmiller, J E Diseases of purine and pyrimidine metabolism. *In* Bondy, P K., and Rosenberg L E (eds.) Metabolic Control and Disease, 8th Ed. Philadelphia, W B Saunders Co, 1980, p 777

Sugden, P H. The effects of hormonal factors on cardiac protein turnover *Adv Myocardiol*, *5* 105 1985

Visek, W J. An update of concepts of essential amino acids. *Annu Rev Neurosci*, *5* 137, 1982

Waterlow, J C (ed) Nitrogen Metabolism in Man New York, Elsevier Science Publishing Co, 1982.

Waterlow, J C., *et al* Protein Turnover in Mammalian Tissues and in the Whole Body New York, Elsevier/North Holland, 1978

Zak, R., *et al* Assessment of protein turnover by use of radioisotope tracers *Physiol. Rev.*, *59* 407, 1979

Zamyatnin, A A. Amino acid, peptide, and protein volume in solution *Annu. Rev Biophys Bioeng*, *13* 145, 1984

70

The Liver as an Organ

In many previous chapters we have encountered the liver, and each time it is performing a different function. Yet, we often forget that the liver is a discrete organ and that many of its various functions interrelate to each other. This becomes particularly evident in clinical abnormalities of the liver because many of its functions are disturbed simultaneously but in different combinations, depending upon the nature of the disorder. Therefore, the purpose of the present chapter is to summarize the different functions of the liver and to show how the liver operates as an individual organ.

The basic functions of the liver can be divided into (1) its vascular functions for storage and filtration of blood, (2) its metabolic functions concerned with the majority of the metabolic systems of the body, and (3) its secretory and excretory functions that are responsible for forming the bile that flows through the bile ducts into the gastrointestinal tract.

PHYSIOLOGIC ANATOMY OF THE LIVER

The basic functional unit of the liver is the *liver lobule*, which is a cylindrical structure several millimeters in length and 0.8 to 2 mm in diameter. The human liver contains 50,000 to 100,000 individual lobules.

The liver lobule is constructed around a *central vein* that empties into the hepatic veins and thence into the vena cava. The lobule itself is composed principally of many *hepatic cellular plates* (two of which are illustrated in Figure 70–1) that radiate centrifugally from the central vein like spokes in a wheel. Each hepatic plate is one to two cells thick, and between the adjacent cells lie small *bile canaliculi* that empty into *bile ducts* that originate in the fibrous septa that separate the adjacent liver lobules.

Also in the septa are small *portal venules* that receive their blood from the portal veins. From these venules blood flows into flat, branching *hepatic sinusoids* that lie between the hepatic plates, and thence into the central vein. Thus, the hepatic cells are exposed continuously to portal venous blood.

In addition to the portal venules, *hepatic arterioles* are also present in the interlobular septa. These arterioles supply arterial blood to the septal tissues, and many of the small arterioles also empty directly into the hepatic sinusoids, most frequently emptying into these about one third of the distance away from the interlobular septa, as shown in Figure 70–1.

The venous sinusoids are lined by two types of cells: (1) typical *endothelial cells* and (2) large *Kupffer cells*, which are tissue macrophages (also called reticuloendothelial cells) capable of phagocytizing bacteria and other foreign matter in the blood. The endothelial lining of the venous sinusoids has extremely large pores, some of which are almost 1 micron in diameter. Beneath this lining, lying between the endothelial cells and the hepatic cells, is a very narrow tissue space called the *space of Disse*. Because of the large pores in the endothelium, substances in the plasma move freely into the space of

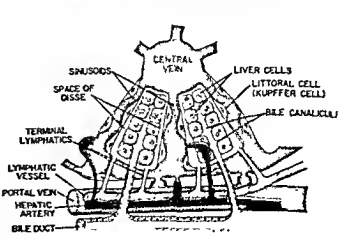


Figure 70–1. Basic structure of a liver lobule showing the hepatic cellular plates, the blood vessels, the bile-collecting system, and the lymph flow system comprised of the spaces of Disse and the interlobular lymphatics. (Reprinted from Guyton, Taylor, and Granger [as modified from Elias]: Dynamics of the Body Fluids, 1975.)

Disse. Even large portions of the plasma proteins diffuse freely into this space.

In the interlobular septa are vast numbers of *terminal lymphatics* The spaces of Disse connect directly with the lymphatics so that excess fluid in these spaces is removed through the lymphatics

FUNCTION OF THE HEPATIC VASCULAR SYSTEM

The function of the hepatic vascular system was discussed in Chapter 29 in connection with the portal veins Briefly, this can be summarized as follows·

Blood Flow Through the Liver. About 1100 ml of blood flows from the portal vein into the liver sinusoids each minute, and approximately an additional 350 ml flows into the sinusoids from the hepatic artery, the total averaging about 1450 ml per minute.

Total hepatic blood flow per minute can be measured by a modified Fick procedure in which the dye *indocyanine green*, a dye that is removed from the blood only by the liver, is infused continuously into the circulatory system until its concentration in the arterial blood remains constant When this is achieved, one knows the rate of excretion of the dye because it is then exactly equal to the rate of infusion Next, the concentration of the dye is measured in the arterial blood and also in venous blood collected from the hepatic vein by means of a catheter From these measurements, one can calculate the arteriovenous difference of the dye The liver blood flow is then calculated by the usual Fick formula, as follows

$$\text{Hepatic blood flow} = \frac{\text{Rate of dye excretion}}{\text{A-V difference in dye}}$$

Pressures and Resistance in the Hepatic Vessels. The pressure in the hepatic vein leading from the liver into the vena cava averages almost exactly 0 mm Hg, whereas the pressure in the portal vein leading into the liver averages 9 mm Hg This shows that the resistance to blood flow through the liver sinusoids is normally low, especially when one considers that about 1.45 liters of blood flow by this route each minute However, various pathological conditions can cause the resistance to rise markedly, sometimes increasing the portal venous pressure to as high as 20 to 30 mm Hg The most common cause of increased hepatic vascular resistance is the disease *liver cirrhosis*, in which many of the vascular channels through the liver are severely restricted by fibrotic constriction of the sinusoids or even complete blockage or destruction

Storage of Blood in the Liver; Hepatic Congestion. An increase in pressure in the veins draining the liver dams blood in the liver sinusoids and thereby causes the entire liver to swell markedly The liver can store 200 to 400 ml of blood in this way as the result of only a 4 to 8 mm Hg rise in hepatic venous pressure For this reason, the liver is one of the major *blood reservoirs* Conversely, if a person hemorrhages so that large amounts of blood are lost from the circulatory system, much of the normal blood in the liver sinusoids drains into the remainder of the circulation to help replace the lost blood.

The most common cause of hepatic congestion is cardiac failure, which often increases the central venous pressure to as high as 10 to 15 mm Hg. The continual stretching of the liver sinusoids that results and the stasis of blood caused by the hepatic congestion gradually lead to necrosis of many of the hepatic cells in the hepatic cellular plates.

Lymph Flow from the Liver

Because the pores in the hepatic sinusoids allow ready passage of proteins into the spaces of Disse, the lymph draining from the liver usually has a protein concentration of about 6 grams per deciliter, which is only slightly less than the protein concentration of plasma. Also, the extreme permeability of the liver sinusoids allows large quantities of lymph to form Therefore, about one half of all the lymph formed in the body under resting conditions arises in the liver.

Effects of High Hepatic Vascular Pressures on Fluid Transudation from the Liver Sinusoids and Portal Capillaries. When the hepatic venous pressure rises only 3 to 7 mm Hg above normal, excessive amounts of fluid begin to transude into the lymph and also to leak through the outer surface of the liver capsule directly into the abdominal cavity. This fluid is almost pure plasma, containing 80 to 90 per cent as much protein as normal plasma At still higher hepatic venous pressures, at 10 to 15 mm Hg, liver lymph flow increases to as much as 20 times normal, and the "sweating" from the surface of the liver can be so great that it causes severe *ascites* (free fluid in the abdominal cavity).

Blockage of portal flow into or through the liver causes very high capillary pressures in the entire gastrointestinal tract, resulting in edema of the gut wall and transudation of fluid through the serosa of the gut into the abdominal cavity This, too, can cause ascites but is less likely to do so than is sweating from the liver surface, because collateral vascular channels develop rapidly from the portal veins to the systemic veins, decreasing the intestinal capillary pressure back to a safe value

The Hepatic Tissue Macrophage System

The inner surfaces of all the liver sinusoids are loaded with many *Kupffer cells*, which protrude into the flowing blood as illustrated in Figure 70–1 and as shown in detail in Figure 5–4 of Chapter 5. These cells are highly phagocytic, so much so that they can remove 99 per cent (or more) of bacteria in the portal venous blood before they can pass all the way through the liver sinusoids Since the portal blood drains from the intestines, it almost always contains a considerable number of colon bacilli Therefore, the importance of the Kupffer cell filtration system is readily apparent The number of Kupffer cells in the sinusoids increases markedly when increased quantities of particulate matter or other debris are present in the blood.

METABOLIC FUNCTIONS OF THE LIVER

The liver cells all together are a large chemically reactant pool having a very high rate of metabolism,

sharing substrates and energy from one metabolic system to another, processing and synthesizing multiple substances that are transported to other areas of the body, and performing a myriad of other metabolic functions. For all of these reasons, a major share of the entire discipline of biochemistry is devoted to the metabolic reactions in the liver. But, here, let us summarize those metabolic functions that are especially important in understanding the integrated physiology of the body.

CARBOHYDRATE METABOLISM

In carbohydrate metabolism the liver performs the following specific functions. (1) storage of glycogen, (2) conversion of galactose and fructose to glucose, (3) gluconeogenesis, and (4) formation of many important chemical compounds from the intermediate products of carbohydrate metabolism.

The liver is especially important for maintaining a normal blood glucose concentration. For instance, storage of glycogen allows the liver to remove excess glucose from the blood, store it, and then return it to the blood when the blood glucose concentration begins to fall too low. This is called the *glucose buffer function* of the liver. As an example, immediately after a meal containing large amounts of carbohydrates, the blood glucose concentration rises about three times as much in a person with a nonfunctional liver as in a person with a normal liver.

Gluconeogenesis in the liver is also concerned with maintaining a normal blood glucose concentration, for gluconeogenesis occurs to a significant extent only when the glucose concentration begins to fall below normal. In such a case, large amounts of amino acids are converted into glucose, thereby contributing still another means for maintaining a relatively normal blood glucose concentration.

FAT METABOLISM

Though some fat metabolism can take place in almost all cells of the body, certain aspects of fat metabolism occur mainly in the liver. Some specific functions of the liver in fat metabolism are (1) very high rate of beta oxidation of fatty acids and formation of acetoacetic acid, (2) formation of most of the lipoproteins, (3) synthesis of large quantities of cholesterol and phospholipids, and (4) conversion of large quantities of carbohydrates and proteins to fat.

To derive energy from neutral fats, the fat is first split into glycerol and fatty acids; then the fatty acids are split by *beta oxidation* into two-carbon acetyl radicals that then form *acetylcoenzyme A* (acetyl·CoA). This in turn can enter the citric acid cycle and be oxidized to liberate tremendous amounts of energy. Beta oxidation can take place in all cells of the body, but it occurs especially rapidly in the hepatic cells. Yet the liver itself cannot utilize all the acetyl·CoA that is formed; instead, this is converted by condensation of two molecules of acetyl·CoA into *acetoacetic acid*, which is a highly soluble acid that passes from the liver cells into the extracellular fluids and then is transported throughout the body to be absorbed by the other tissues. These tissues in turn reconvert the acetoacetic acid into acetyl·CoA and then oxidize it in the usual manner. In this way, therefore, the liver is responsible for a major part of the metabolism of fats.

About 80 per cent of the cholesterol synthesized is converted into bile salts, which in turn are secreted into the bile; the remainder is transported in the lipoproteins, which are carried by the blood to the tissue cells everywhere in the body. Phospholipids are likewise synthesized in the liver and transported principally in the lipoproteins. Both the cholesterol and the phospholipids are used by the cells to form membranes, intracellular structures, and multiple derived chemical substances that are important to cellular function.

Most of the fat synthesis in the body from carbohydrates and proteins also occurs in the liver. After fat is synthesized in the liver it is transported in the lipoproteins to the adipose tissue to be stored.

PROTEIN METABOLISM

Even though a large proportion of the processes for carbohydrate and fat metabolism occurs in the liver, the body could probably dispense with many of these functions of the liver and still survive. On the other hand, the body cannot dispense with the services of the liver in protein metabolism for more than a few days without death ensuing. The most important functions of the liver in protein metabolism are (1) deamination of amino acids, (2) formation of urea for removal of ammonia from the body fluids, (3) formation of plasma proteins, and (4) interconversions among the different amino acids and other compounds important to the metabolic processes of the body.

Deamination of the amino acids is required before these can be used for energy or before they can be converted into carbohydrates or fats. A small amount of deamination can occur in the other tissues of the body, especially in the kidney, but the percentage of deamination occurring extrahepatically is so small that it is almost completely unimportant.

Formation of urea by the liver removes ammonia from the body fluids. Large amounts of ammonia are formed by the deamination process, and still additional amounts are continually formed in the gut by bacteria and are then absorbed into the blood. Therefore, in the absence of this function of the liver to form urea, the plasma ammonia concentration rises rapidly and results in *hepatic coma* and death. Indeed, even greatly decreased blood flow through the liver—as occurs occasionally when a shunt develops between the portal vein and the vena cava—can also cause excessive ammonia in the blood, an exceedingly toxic condition.

Essentially all the plasma proteins, with the exception of part of the gamma globulins, are formed by the hepatic cells. This accounts for about 90 per cent of all the plasma proteins. The remaining gamma globulins are the antibodies formed mainly by plasma cells in the lymph tissue of the body. The liver can probably form plasma proteins at a maximum rate of 15 to 50 grams per day. Therefore, after loss of as much as half the plasma proteins from the body, these can be replenished in a week or two. It is particularly interesting that plasma protein depletion causes rapid mitosis of the hepatic cells and actual growth of the liver to a larger size; these effects are coupled with rapid output of plasma proteins until the plasma concentration returns to normal.

Among the most important functions of the liver is its ability to synthesize certain amino acids and also to synthesize other important chemical compounds from amino acids. For instance, the so-called nonessential amino acids can all be synthesized in the liver. To do this, a keto acid having the same chemical composition (except at the keto oxygen) as that of the amino acid to be formed is first synthesized. Then an amino radical is transferred through several stages of *transamination* from an available amino acid to the keto acid to take the place of the keto oxygen.

MISCELLANEOUS METABOLIC FUNCTIONS OF THE LIVER

Storage of Vitamins. The liver has a particular propensity for storing vitamins and has long been known as an excellent source of certain vitamins in treating patients. The single vitamin stored to the greatest extent in the liver is vitamin A, but large quantities of vitamin D and vitamin B_{12} are normally stored as well. Sufficient quantities of vitamin A can be stored to prevent vitamin A deficiency for as long as ten months. Sufficient vitamin D can be stored to prevent deficiency for as long as three to four months, and vitamin B_{12} can be stored for at least a year and maybe several years.

Relationship of the Liver to Blood Coagulation. The liver forms a large proportion of the blood substances utilized in the coagulation process. These are fibrinogen, prothrombin, accelerator globulin, factor VII, and several other important coagulation factors. Vitamin K is required by the metabolic processes of the liver for the formation of prothrombin and factors VII, IX, and X. In the absence of vitamin K the concentrations of these substances fall very low, and this almost prevents blood coagulation.

Storage of Iron. Except for the iron in the hemoglobin of the blood, by far the greater proportion of the iron in the body is usually stored in the liver in the form of *ferritin*. The hepatic cells contain large amounts of a protein called *apoferritin*, which is capable of combining with either small or large quantities of iron. Therefore, when iron is available in the body fluids in extra quantities, it combines with the apoferritin to form ferritin and is stored in this form until needed elsewhere. When the iron in the circulating body fluids reaches a low level, the ferritin releases the iron. Thus, the apoferritin-ferritin system of the liver acts as a *blood iron buffer* and also as an iron storage medium. Other functions of the liver in relation to iron metabolism are considered in Chapter 4.

Removal or Excretion of Drugs, Hormones, and Other Substances by the Liver. The very active chemical medium of the liver is well known for its capability of detoxifying or excreting into the bile many different drugs, including sulfonamides, penicillin, ampicillin, erythromycin, and many others. In a similar manner, several of the different hormones secreted by the endocrine glands are either chemically altered or excreted by the liver, including thyroxine and essentially all the steroid hormones such as estrogen, cortisol, aldosterone, and so forth. Therefore, liver damage can often lead to excess accumulation of one or more of these hormones in the body fluids and therefore can cause overactivity of the hormonal systems. Finally, one of the major routes for excreting calcium from the blood is into the bile and then into the feces.

EXCRETION OF BILIRUBIN IN THE BILE

The formation of bile by the liver and its function in the digestive and absorptive processes of the intestinal tract were discussed in Chapters 64 and 65. In addition, many substances are excreted in the bile and then eliminated in the feces. One of these is the greenish-yellow pigment *bilirubin*. This is a major end-product of hemoglobin degradation, as pointed out in Chapter 4. It is highly soluble in all cell membranes and is also very toxic. Therefore, its excretion in the bile is one of the very important functions of the liver. A full understanding of bilirubin formation and excretion provides an exceedingly important tool to the clinician for diagnosing both hemolytic blood diseases and various types of liver diseases.

Briefly, when the red blood cells have lived out their life span, averaging 120 days, and have become too fragile to exist longer in the circulatory system, their cell membranes rupture and the released hemoglobin is phagocytized by tissue macrophages (also called "reticuloendothelial system") throughout the body. Here, the hemoglobin is first split into *globin* and *heme*, and the heme ring is opened to give (a) free iron that is transported in the blood by transferrin, and (b) a straight chain of four pyrrole nuclei that is the substrate from which the bile pigments are formed. The first pigment formed is *biliverdin*, but this is rapidly reduced to *free bilirubin*, which is gradually released into the plasma. However, the free bilirubin immediately combines very strongly with the plasma albumin and is transported in this combination throughout the blood and interstitial fluids. Even when bound with the plasma protein, this bilirubin is still called "free bilirubin" to distinguish it from "conjugated bilirubin," which will be discussed later. Within hours, the free bilirubin is absorbed through the hepatic cell membrane, in this process being released from the plasma albumin but almost instantly being combined with one of two proteins (called "Y" and "Z" proteins) inside the hepatic cells that trap the bilirubin inside the cells. Soon thereafter, however, the bilirubin is also removed from this protein and conjugated with other substances. About 80 per cent of it conjugates with glucuronic acid to form bilirubin *glucuronide*; an additional 10 per cent conjugates with sulfate to form *bilirubin sulfate*, and the final 10 per cent conjugates with a multitude of other substances. It is in these forms that the bilirubin is excreted by an active transport process into the bile canaliculi.

A small portion of the *conjugated bilirubin* formed by the hepatic cells returns to the plasma, either directly into the liver sinusoids or indirectly by absorption into the blood from the bile ducts or lymphatics. Regardless of the exact mechanism by which it re-enters the blood, this causes a small portion of the bilirubin in the extracellular fluids always to be of the conjugated type rather than of the free type.

Formation and Fate of Urobilinogen. Once in the intestine, about one half of the bilirubin is converted by bacterial action into the substance *urobilinogen*, which is highly soluble. Some of the urobilinogen is reabsorbed through the intestinal mucosa into the blood. Most of this is re-excreted by the liver back into the gut, but about 5 per cent is excreted by the kidneys into the urine. After exposure to air in the urine, the urobilinogen becomes oxidized to *urobilin*, or in the feces it

becomes altered and oxidized to form *stercobilin*. These terrelationships of bilirubin and the other bile pigents are illustrated in Figure 70–2.

Jaundice. The word "jaundice" means a yellowish nt to the body tissues, including yellowness of the skin d also of the deep tissues. The usual cause of jaundice large quantities of bilirubin in the extracellular fluids, ther free bilirubin or conjugated bilirubin. The normal asma concentration of bilirubin, including both the ee and the conjugated forms, averages 0.5 mg/dl of lasma. However, in certain abnormal conditions this in rise to as high as 40 mg/dl The skin usually begins appear jaundiced when the concentration rises to bout three times normal—that is, above 1.5 mg/dl

The common causes of jaundice are (1) increased estruction of red blood cells with rapid release of ilirubin into the blood or (2) obstruction of the bile ucts or damage to the liver cells so that even the usual mounts of bilirubin cannot be excreted into the gas-ointestinal tract. These two types of jaundice are alled, respectively, *hemolytic jaundice* and *obstructive undice*. They differ from each other in the following ays:

Hemolytic Jaundice. In hemolytic jaundice, the excretory function of the liver is not impaired in the least, ut red blood cells are hemolyzed rapidly and the hepatic ells simply cannot excrete the bilirubin as rapidly as it formed. Therefore, the plasma concentrations of both free and conjugated bilirubin rise to levels much above normal. Likewise, the rate of formation of *urobilinogen* in the intestine is greatly increased, and much of this is absorbed into the blood and later excreted in the urine.

Obstructive Jaundice. In obstructive jaundice, caused either by obstruction of the bile ducts or by damage to the liver cells, the rate of bilirubin formation is normal, but the bilirubin formed simply cannot pass from the blood into the intestines. However, the free bilirubin usually does still enter the liver cells and becomes conjugated in the usual way. This conjugated bilirubin is then returned to the blood probably by rupture of the congested bile canaliculi and direct emptying of the bile into the lymph leaving the liver. Thus, *most of the bilirubin in the plasma becomes the conjugated type* rather than the free type

Diagnostic Differences Between Hemolytic and Obstructive Jaundice. A simple test called the *van den Bergh test* can be used to differentiate between free and conjugated bilirubin in the plasma. If an immediate reaction occurs with the van den Bergh reagent (which gives a colorimetric change), the bilirubin is of the conjugated type, and the reaction is called a *direct* van den Bergh reaction. However, to demonstrate the presence of free bilirubin, one must add alcohol to the plasma. This precipitates the protein and "frees" the "free" bilirubin from its protein complex so that it can

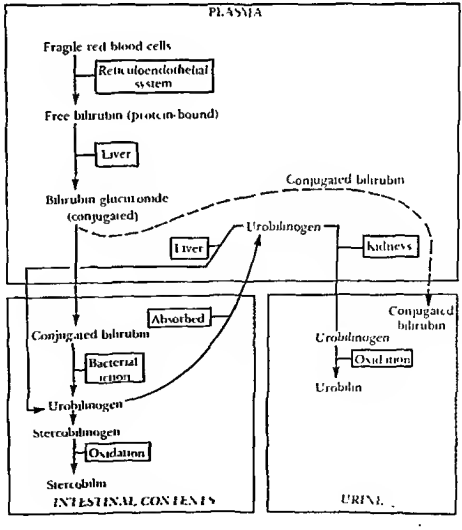


Figure 70–2. Bilirubin formation and excretion.

then combine with the van den Bergh reagent This causes the colorimetric change to be much stronger, and the additional result is called the *indirect* van den Bergh reaction Thus, *in hemolytic jaundice an indirect van den Bergh reaction occurs (increased free bilirubin) and in obstructive jaundice a direct van den Bergh reaction occurs (increased conjugated bilirubin)*

When there is total obstruction of bile flow, no bilirubin at all can reach the intestines to be converted into urobilinogen by bacteria Therefore, urobilinogen is not reabsorbed into the blood and also is not excreted by the kidneys into the urine Consequently, in *total* obstructive jaundice, tests for urobilinogen in the urine are completely negative Also, the stools become clay colored for lack of stercobilin and other bile pigments

Another major difference between free and conjugated bilirubin is that kidneys can excrete the highly soluble conjugated bilirubin but not the albumin-bound "free" bilirubin Therefore, in severe obstructive jaundice large quantities of conjugated bilirubin appear in the urine This can be demonstrated simply by shaking the urine and observing the foam, which becomes colored an intense yellow

Thus, by understanding the physiology of bilirubin excretion by the liver and by use of a few simple tests, it is possible to differentiate between obstructive and hemolytic jaundice and, often, also to determine the severity of the disease

REFERENCES

Alagille, D and Odievre M Liver and Biliary Tract Disease in Children New York, John Wiley & Sons, 1979

Berk P D et al Disorders of bilirubin metabolism *In* Bondy, P K and Rosenberg L E (eds) Metabolic Control and Disease 8th Ed Philadelphia W B Saunders Co 1980 p 1009

Boucher I A D The medical treatment of gallstones. *Annu Rev Med 31* 59 1980

Boyer J L New concepts of mechanisms of hepatocyte bile formation *Physiol Rev 60* 303 1980

Brown H and Hardwick, D F (eds) Intermediary Metabolism of the Liver Springfield Ill Charles C Thomas, 1973

Davidson C S (ed) Problems in Liver Diseases. New York, Stratton Intercontinental Medical Book Corp 1979

El Maghrabi M R and Pilkis, S J Rat liver 6-phosphofructo-2 kinase fructose 2 6 biophosphatase A review of relationships between the two activities of the enzyme *J Cell Biochem 26* 1, 1984

Erlinger S and Dhumeaux, D Mechanisms and control of secretion of bile water and electrolytes *Gastroenterology, 66* 281, 1974

Falcon W W Hepatic mechanisms *In* Frohlich, E D (ed) Pathophysiology 2nd Ed Philadelphia, J B Lippincott Co, 1976 p 531

Farber E and Fisher, M M (eds) Toxic Injury of the Liver New York Marcel Dekker, 1979

Fevery J and Heirwegh K P M Bilirubin metabolism *In* Javitt, N B (ed) International Review of Physiology Liver and Biliary Tract Physiology I Vol 21 Baltimore, University Park Press, 1980, p 171

Fisher, M M *et al* (eds) Gall Stones New York, Plenum Press 1979

Forker, E L Mechanisms of hepatic bile formation *Annu Rev Physiol 39* 323, 1977.

Fortie T M Primary hepatocytes in monolayer culture A model for studies on lipoprotein metabolism *Annu Rev Physiol, 46* 403 19[illegible]

Frizzell R A, and Heintze K Transport functions of the gallbladder *In* Javitt, N B (ed) International Review of Physiology Liver and Biliary Tract Physiology I Vol 21, Baltimore, University Park Press 1980, p 221

Gitlambos, J T Cirrhosis Philadelphia, W B Saunders Co, 1979

Gall E A and Mostofi, F K (eds) The Liver Huntington, N Y, E Krieger Publishing Co, 1980

Gerolami, A, and Sarles, J C Biliary secretion and motility *Int Rev Physiol, 12* 223, 1977

Gitnick, G L (ed) Current Gastroenterology and Hepatology Boston Houghton Mifflin, 1979

Gorosky, C A, and Groom, A C Microcirculatory events in the liver and the spleen *In* Renkin, E M, and Michel, C C (eds) Handbook of Physiology Sec 2, Vol IV Bethesda, American Physiological Society, 1984, p 689

Gustafsson, J A Sex steroid induced changes in hepatic enzymes *Annu Rev Physiol 45* 51, 1983

Henderson, J R, and Daniel, P M Capillary beds and portal circulations. *In* Renkin, E M, and Michel, C C (eds) Handbook of Physiology Sec 2, Vol IV Bethesda American Physiological Society 1984 p 1035

Jones R S, and Myers, W C Regulation of hepatic biliary secretion. *Annu Rev Physiol, 41* 67, 1979

Kappas A, and Alvares, A P How the liver metabolizes foreign substances *Sci Am, 232*(6) 22, 1975

Kraus Friedmann N Hormonal regulation of hepatic gluconeogenesis. *Physiol Rev, 64* 170, 1984

Langee, M The Endocrine and the Liver New York, Academic Press, 1983

LaRusso, N F Proteins in bile How they get there and what they do *Am J Physiol, 247* G199 1984

MacSween R N M, *et al* (eds) Pathology of the Liver New York Longman, Inc, 1979

McGarry J D, and Foster, D W Regulation of hepatic ketogenesis *In* DeGroot L J *et al* (eds) Endocrinology Vol 2 New York, Grune & Stratton, 1979 p 997

Paumgartner G, *et al* (eds) Biological Effects of Bile Acids Baltimore, University Park Press, 1979

Presig, R, *et al* Physiologic and pathophysiologic aspects of the hepatic hemodynamics *Prog Liver Dis 4* 201, 1972

Rappaport, A M Hepatic blood flow Morphologic aspects and physiologic regulation *In* Javitt, N B (ed) International Review of Physiology Liver and Biliary Tract Physiology I Vol 21 Baltimore University Park Press, 1980, p 1

Reichen, J, and Paumgartner, G Excretory function of the liver *In* Javitt N B (ed) International Review of Physiology Liver and Biliary Tract Physiology I Vol 21 Baltimore, University Park Press, 1980, p 103

Richardson, P D I, and Withrington, P G Physiological regulation of the hepatic circulation *Annu Rev Physiol, 44* 57, 1982

Salen, G and Shefer, S Bile acid synthesis *Annu Rev Physiol, 45* 679 1983

Schachter, D Fluidity and function of hepatocyte plasma membranes. *Hepatology 4* 140 1984

Shearman, D J (ed) Diseases of the Gastrointestinal Tract and Liver New York Churchill Livingstone, 1982

Sherlock, S The Human Liver Burlington, N C, Carolina Biological Supply Co, 1978

Soloway, R D (ed) Chronic Active Liver Disease New York, Churchill Livingstone, 1983

Strange, R C Hepatic bile flow *Physiol Rev, 64* 1055, 1984

Tanikawa, K Ultrastructural Aspects of the Liver and Its Disorders New York Igaku-Shoin, 1979

71

Energetics and Metabolic Rate

IMPORTANCE OF ADENOSINE TRIPHOSPHATE (ATP) IN METABOLISM

In tha last few chapters it has been pointed out that carbohydrates, fats, and proteins can all be used by cells to synthesize large quantities of ATP, and that in turn the ATP can be used as an energy source for many other cellular functions For thesa reasons, ATP has been called an energy "currency" that can be created and expended. Indeed, tha cells can transfer energy from the different foodstuffs to most functional systems of the cells only through this medium of ATP (or the very similar nucleotida guanosina triphosphate (GTP)) Many of the attributes of ATP were presented in Chapter 2, but others require discussion at this point

An attributa of ATP that makes it higbly valuable as a means of energy currency is the larga quantity of free energy (about 7300 calories per mole under standard conditions but as much as 12,000 calories under physiological conditions) vested in each of its two high energy phosphate bonds. The amount of energy in each bond, when liberated by decomposition of ATP, is enough to cause almost any step of any chemical reaction in the body to take place if appropriate transfer of the energy is achieved. Some chemical reactions that require ATP energy use only a few hundred of the available 12,000 calories, and the remainder of this energy is then lost in the form of heat. Yet even this inefficiency in the utilization of energy is better than not being able to energize the necessary chemical reactions at all

The precise methods by which ATP is used to cause the many physical, chemical, and other types of functions in the cells lie principally in the province of biochemistry. Therefore, only the more important functions of ATP are listed here.

Use of ATP for Synthesis of Important Cellular Components. Among the most important of the intracellular processes that require ATP is the formation of peptide linkages between amino acids during the synthesis of proteins. The different peptide linkages, depending on which types of amino acids are linked together, require from 500 to 4000 calories of energy per mole. It will be recalled from the discussion of protein synthesis in Chapter 3 that four separate high energy phosphate bonds are expended during the cascade of reactions required for forming each peptide linkage. This provides a total of 48,000 calories of energy, which is far more than the 500 to 5000 calories eventually stored in each of the different peptide linkages.

Also, it will be recalled from the preceding chapters that ATP is utilized in the synthesis of glucose from lactic acid and in the synthesis of fatty acids from acetyl Co-A In addition, ATP is utilized for synthesis of cholesterol, phospholipids, the hormones, and almost all other substances of the body. Even the urea excreted by the kidneys requires ATP to cause its formation. One might wonder at the advisability of expending energy to form urea which then is simply thrown away from the body However, remembering the extreme toxicity of ammonia in the body fluids, one can see the value of this reaction, which keeps the ammonia concentration of tha body fluids always at a low level

Use of ATP for Muscular Contraction. Muscular contraction will not occur without energy from ATP Myosin, one of the important contractila proteins of the muscle fiber, acts as an enzyme to cause breakdown of ATP into adenosine diphosphate (ADP), thus causing the release of energy Only a small amount of ATP is normally degraded in muscles when muscular contraction is not occurring, but this rate of ATP usage can rise to at least 150 times the resting level during short bursts of maximal contraction The postulated mechanism by which ATP is utilized to cause muscle contraction was discussed in Chapter 11

Use of ATP for Active Transport Across Membranes. In Chapters 9, 34, and 65, active transport of electrolytes and various nutrients across cell membranes and from the renal tubules and gastrointestinal tract was discussed. In each instance, it was noted that active transport of most electrolytes and other substances such as glucose, amino acids, and acetoacetate can occur against an electrochemical gradient, even though the natural diffusion of the substances would be in the opposite direction. Obviously, to oppose the electrochemical gradient requires energy, as was discussed in Chapter 9 This energy is provided by ATP

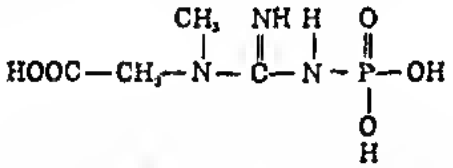
Energy for Glandular Secretion. The same principles apply to glandular secretion as to the absorption of substances against concentration gradients, for energy is required to concentrate substances as they are secreted by the glandular cell. In addition, energy is also required to synthesize the organic compounds to be secreted.

Energy for Nerve Conduction. The energy utilized during propagation of a nerve impulse is derived from the potential energy stored in the form of concentration differences of ions across the membranes. That is, a high concentration of potassium inside the fiber and a low concentration outside the fiber constitutes a type of energy storage. Likewise, a high concentration of sodium

on the outside of the membrane and a low concentration on the inside represents another store of energy. The energy needed to pass each action potential along the fiber membrane is derived from this energy storage, with small amounts of potassium transferring out of the cell and sodium into the cell during each of the action potentials. However, active transport systems then retransport the ions back through the membrane to their former positions. Here, ATP is utilized in abundance to retransfer the sodium and potassium ions after nerve action potentials have been conducted, and the rate of ATP usage increases in proportion to the number of action potentials transmitted.

PHOSPHOCREATINE AS A STORAGE DEPOT FOR ENERGY AND FOR BUFFERING THE CONCENTRATION OF ATP

Despite the paramount importance of ATP as a coupling agent for energy transfer, this substance is not the most abundant store of high energy phosphate bonds in the cells. On the contrary, *phosphocreatine*, which also contains high energy phosphate bonds, is three to eight times as abundant. The high energy bond of phosphocreatine contains about 8500 calories per mole under standard conditions, and as much as 13,000 calories per mole under conditions in the body (38° C and low concentrations of the reactants). This is somewhat greater than the 12,000 calories per mole in each of the two high energy phosphate bonds of ATP. The formula for creatine phosphate is the following:



Phosphocreatine cannot act in the same manner as ATP as a coupling agent for transfer of energy between the foods and the functional cellular systems. But it can transfer energy interchangeably with ATP. When extra amounts of ATP are available in the cell, much of its energy is utilized to synthesize phosphocreatine, thus building up this storehouse of energy. Then when the ATP begins to be used up, the energy in the phosphocreatine is transferred rapidly back to ATP and from this to the functional systems of the cells. This reversible interrelationship between ATP and creatine phosphate is illustrated by the following equation:

$$\text{Phosphocreatine} + \text{ADP} \rightleftarrows \text{ATP} + \text{Creatine}$$

Note particularly that the higher energy level of the high energy phosphate bond in phosphocreatine—1000 to 1500 calories per mole greater than that in ATP—causes the reaction between phosphocreatine and ATP to proceed to an equilibrium state about 25 times in favor of ATP. Therefore, the slightest utilization of ATP by the cells calls forth the energy from the phosphocreatine to synthesize new ATP. This effect keeps the concentration of ATP at an almost constant level as long as any phosphocreatine remains. For this reason we can call the ATP-phosphocreatine system an ATP "buffer" system. One can readily understand the importance of keeping the concentration of ATP very nearly constant, because the rates of almost all the reactions in the body depend on this constancy.

ANAEROBIC VERSUS AEROBIC ENERGY

Anaerobic energy means energy that can be derived from foods without the simultaneous utilization of oxygen; *aerobic energy* means energy that can be derived from foods only by oxidative metabolism. In the discussions in Chapters 67 through 69, it was noted that carbohydrates, fats, and proteins can all be oxidized to cause synthesis of ATP. However, carbohydrates are the only significant foods that can be utilized to provide energy without utilization of oxygen; this energy release occurs during glycolytic breakdown of glucose or glycogen to pyruvic acid. For each mole of glucose that is split into pyruvic acid, 2 moles of ATP are formed. However, when glycogen is split to pyruvic acid, each mole of glucose in the glycogen gives rise to 3 moles of ATP. The reason for this difference is that free glucose entering the cell must be phosphorylated by 1 mole of ATP before it can begin to be split, whereas this is not true of glucose derived from glycogen because it comes from the glycogen already in the phosphorylated state without additional expenditure of ATP. Thus, the best source of energy under anaerobic conditions is the stored glycogen of the cells.

Anaerobic Energy During Hypoxia. One of the prime examples of anaerobic energy utilization occurs in acute hypoxia. When a person stops breathing, there is already a small amount of oxygen stored in the lungs and an additional amount stored in the hemoglobin of the blood. However, these are sufficient to keep the metabolic processes functioning only for about two minutes. Continued life beyond this time requires an additional source of energy. This can be derived for another minute or so from glycolysis, the glycogen of the cells splitting into pyruvic acid and the pyruvic acid in turn becoming lactic acid, which diffuses out of the cells as described in Chapter 67.

Anaerobic Energy Usage in Strenuous Bursts of Activity. It is common knowledge that muscles can perform extreme feats of strength for a few seconds but are much less capable during prolonged activity. The energy used during strenuous activity is derived from (1) ATP already present in the muscle cells, (2) phosphocreatine in the cells, (3) anaerobic energy released by glycolytic breakdown of glycogen to lactic acid, and (4) oxidative energy released continuously by oxidative processes in the cells. The speed of the oxidative processes cannot approach that which is required to supply all the energy demands during strenuous bursts of activity. Therefore, the other three sources of energy are called upon to their maximum extent.

The maximum amount of ATP in muscle is only about 5 millimoles per liter of intracellular fluid, and this amount can maintain maximum muscle contraction for not more than a second or so. The amount of phosphocreatine in the cells is three to eight times this amount, but even by utilization of all the phosphocreatine, the amount of time that maximum contraction can be maintained is still only a few more seconds. Release of energy by glycolysis can occur much more rapidly than can oxidative release of energy. Consequently, most of the extra energy required during strenuous activity that

lasts for more than a few seconds but less than one to two minutes is derived from anaerobic glycolysis. As a result, the glycogen content of muscles after strenuous bouts of exercise becomes reduced, whereas the lactic acid concentration of the blood rises. Then, immediately *after the exercise is over, oxidative metabolism is used* to reconvert about four fifths of the lactic acid into glucose, while the remainder becomes pyruvic acid and is degraded and oxidized in the citric acid cycle. The reconversion to glucose occurs principally in the liver cells, and the glucose is then transported in the blood back to the muscles where it is stored once more in the form of glucogen.

Oxygen Debt. After a period of strenuous exercise, a person continues to breathe hard and to consume excessive amounts of oxygen for a few minutes or up to as long as an hour thereafter. This excess oxygen is used (1) to reconvert the lactic acid that has accumulated during exercise back into glucose, (2) to reconvert the decomposed ATP and creatine phosphate to their original states, (3) to re-establish normal concentrations of oxygen bound with hemoglobin and myoglobin, and (4) to raise the concentration of oxygen in the lungs up to its normal level. This excess consumption of oxygen after the exercise is over is called the *oxygen debt.*

The principle of oxygen debt will be discussed further in Chapter 84 in relation to sports physiology; the ability of a person to build up an oxygen debt is especially important in many types of athletics

SUMMARY OF ENERGY UTILIZATION BY THE CELLS

With the background of the past few chapters and of the preceding discussion, we can now synthesize a composite picture of overall energy utilization by the cells as illustrated in Figure 71–1. This figure shows the anaerobic utilization of glycogen and glucose to form ATP and also the aerobic utilization of compounds derived from carbohydrates, fats, proteins, and other substances for the formation of still additional ATP. In turn, ATP is in reversible equilibrium with phosphocreatine in the cells, and, since larger quantities of phosphocreatine are present in the cell than ATP, much of the stored energy of the cell is in this energy storehouse.

Energy from ATP can be utilized by the different functioning systems of the cells to provide for synthesis and growth, muscular contraction, glandular secretion, nerve impulse conduction, active absorption, and other cellular activities. If greater amounts of energy are called forth for cellular activities than can be provided by oxidative metabolism, the phosphocreatine storehouse is first utilized, and this is followed rapidly by anaerobic breakdown of glycogen. Thus, oxidative metabolism cannot deliver energy to the cells nearly so rapidly as can the anaerobic processes, but in contrast it is quantitatively almost inexhaustible.

CONTROL OF ENERGY RELEASE IN THE CELL

Rate Control of Enzyme-Catalyzed Reactions. Before it is possible to discuss the control of energy release in the cell, it is necessary to consider the basic principles of *rate control* of enzymatically catalyzed chemical reactions, which are the types of reactions that occur almost universally throughout the body.

The mechanism by which an enzyme catalyzes a chemical reaction is for the enzyme first to combine loosely with one of the substrates of the reaction. This alters the bonding forces on the substrate sufficiently that it can then react with other substances. Therefore, the rate of the overall chemical reaction is determined by both the concentration of the enzyme and the concentration of the substrate that binds with the enzyme. The basic equation expressing this concept is:

$$\text{Rate of reaction} = \frac{K_1 \cdot \text{Enzyme} \cdot \text{Substrate}}{K_2 + \text{Substrate}}$$

This is called the *Michaelis-Menten equation* Figure 71–2 illustrates the application of this equation

Role of Enzyme Concentration in the Regulation of Metabolic Reactions. Figure 71–2 illustrates that

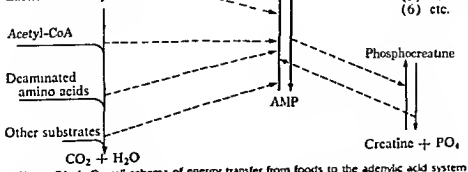


Figure 71–1. Overall schema of energy transfer from foods to the adenylic acid system and then to the functional elements of the cells. (Modified from Soskin and Levine: Carbohydrate Metabolism. University of Chicago Press.)

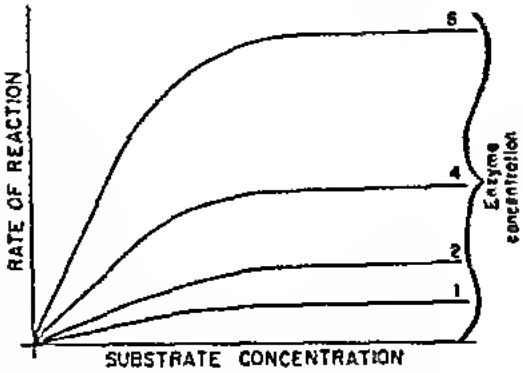


Figure 71–2. Effect of substrate and enzyme concentrations on the rate of enzyme catalyzed reaction

when the substrate is present in reasonably high concentration, as shown in the right half of the figure, the rate of a chemical reaction is then determined almost entirely by the concentration of the enzyme. Thus, as the enzyme concentration increases from an arbitrary value of 1 up to 2, 4, or 8, the rate of the reaction increases proportionately. As an example, when large quantities of glucose enter the renal tubules in diabetes mellitus, the rate of reabsorption of the glucose is determined almost entirely by the concentration of the glucose transport enzymes in the proximal tubular cells because the substrate is then present in excess.

Role of Substrate Concentration in Regulation of Metabolic Reactions. Note also in Figure 71–2 that when the substrate concentration becomes low enough that only a small portion of the enzyme is required in the reaction, the rate of the reaction then becomes directly proportional to the substrate concentration as well as enzyme concentration. This is the effect seen in the absorption of substances from the intestinal tract and renal tubules when their concentrations have become very low.

Rate Limitation in a Series of Reactions. It has become evident from the discussions in the preceding chapters that almost all chemical reactions of the body occur in series, the product of one reaction acting as a substrate of the next reaction, and so on. It is immediately obvious also that the overall rate of a complex series of chemical reactions is determined by the rate of reaction of the slowest step in the series. This is called the *rate-limiting reaction* for the entire series.

Adenosine Diphosphate (ADP) Concentration as a Rate-Controlling Factor in Energy Release. Under *resting* conditions, the concentration of ADP in the cells is extremely slight so that the chemical reactions that depend on ADP as one of the substrates likewise are very slow. These include all the oxidative metabolic pathways as well as essentially all other pathways for release of energy in the body. Thus, *ADP is a major rate-limiting factor* for almost all energy metabolism of the body.

When the cells become active, regardless of the type of activity, ATP is converted into ADP, increasing the concentration of ADP in direct proportion to the degree of activity of the cell. This automatically increases the rates of reactions of the ADP-rate-limiting steps in the metabolic release of energy. Thus, by this simple process, the amount of energy released in the cell is controlled by the degree of activity of the cell. In the absence of cellular activity, the release of energy stops.

THE METABOLIC RATE

The *metabolism* of the body means simply all the chemical reactions in all the cells of the body, and the *metabolic rate* is normally expressed in terms of the rate of heat liberation during the chemical reactions.

Heat as the Common Denominator of All the Energy Released in the Body. In discussing many of the metabolic reactions of the preceding chapters, we have noted that not all the energy in the foods is transferred to ATP, instead, a large portion of this energy becomes heat. On the average, about 50 per cent of the energy in the foods becomes heat during ATP formation. Then still more energy becomes heat as it is transferred from ATP to the functional systems of the cells, so that even under the best of conditions not more than 20 to 25 per cent of all the energy from the food is finally utilized by the functional systems.

Even though 25 per cent of the energy should reach the functional systems of the cells, almost all of this also eventually becomes heat for the following reasons: We might first consider the synthesis of protein and other growing elements of the body. When proteins are synthesized, large portions of ATP are used to form the peptide linkages, and this stores energy in these linkages. But we also noted in our discussions of proteins in Chapter 69 that there is continuous turnover of proteins, some being degraded while others are being formed. When the proteins are degraded, the energy stored in the peptide linkages is released in the form of heat into the body.

Now let us consider the energy used for muscle activity. Much of this energy simply overcomes the viscosity of the muscles themselves or of the tissues so that the limbs can move. The viscous movement in turn causes friction within the tissues, which generates heat.

We might also consider the energy expended by the heart in pumping blood. The blood distends the arterial system, the distention in itself representing a reservoir of potential energy. However, as the blood flows through the peripheral vessels, the friction of the different layers of blood flowing over each other and the friction of the blood against the walls of the vessels turns this energy into heat.

Therefore, essentially all the energy expended by the body is eventually converted into heat. The only significant exception to this occurs when the muscles are used to perform some form of work outside the body. For instance, when the muscles elevate an object to a height or carry the person's body up steps, a type of potential energy is thus created by raising a mass against gravity. But, when external expenditure of energy is not taking place, it is safe to consider that all the energy released by the metabolic processes eventually becomes body heat.

The Calorie. To discuss the metabolic rate and related subjects intelligently, it is necessary to use some unit for expressing the quantity of energy released from the different foods or expended by the different functional processes of the body. Most often, the *Calorie* is the unit used for this purpose. It will be recalled that 1 *calorie*, spelled with a small "c" and often called the *gram-calorie*, is the quantity of heat required to raise the temperature of 1 gram of water 1° C. The calorie is much too small a unit for ease of expression in speaking of energy in the body. Consequently the large Calorie, spelled with a capital "C" and often called the *kilogram-*

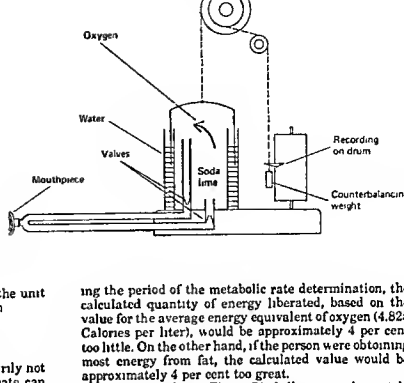


Figure 71–3. The metabolator

calorie, which is equivalent to 1000 calories, is the unit ordinarily used in discussing energy metabolism

MEASUREMENT OF THE METABOLIC RATE

Direct Calorimetry. Since a person is ordinarily not performing any external work, the metabolic rate can be determined by simply measuring the total quantity of heat liberated from the body in a given time This method is called *direct calorimetry*

In determining the metabolic rate by direct calorimetry, one measures the quantity of heat liberated from the body in a large, specially constructed *calorimeter* The subject is placed in an air chamber that is so well insulated that no heat can leak through the walls of the chamber. Heat formed by the subject's body warms the air of the chamber. However, the air temperature within the chamber is maintained at a constant level by forcing the air through pipes in a cool water bath. The rate of heat gain by the water bath, which can be measured with an accurate thermometer, is equal to the rate at which heat is liberated by the subject's body.

Obviously, direct calorimetry is physically difficult to perform and, therefore, is used only for research purposes.

Indirect Calorimetry—The "Energy Equivalent" of Oxygen. Since more than 95 per cent of the energy expended in the body is derived from reactions of oxygen with the different foods, the metabolic rate can also be calculated with a high degree of accuracy from the rate of oxygen utilization. When 1 liter of oxygen is metabolized with glucose, 5.01 Calories of energy are released; when metabolized with starches, 5.06 Calories are released; with fat, 4.70 Calories, and with protein, 4 60 Calories.

From these figures it is striking how nearly equivalent are the quantities of energy liberated per liter of oxygen regardless of the type of food that is being metabolized. For the average diet, the *quantity of energy liberated per liter of oxygen utilized in the body averages approximately 4.825 Calories.* This is called the *energy equivalent* of oxygen; using this energy equivalent, one can calculate with a high degree of precision the rate of heat liberation in the body from the quantity of oxygen utilized in a given period of time.

If a person should metabolize only carbohydrates during the period of the metabolic rate determination, the calculated quantity of energy liberated, based on the value for the average energy equivalent of oxygen (4.825 Calories per liter), would be approximately 4 per cent too little. On the other hand, if the person were obtaining most energy from fat, the calculated value would be approximately 4 per cent too great.

The Metabolator. Figure 71–3 illustrates the metabolator usually used for indirect calorimetry. This apparatus contains a floating drum, under which is an oxygen chamber connected to a mouthpiece through two rubber tubes. A valve in one of these rubber tubes allows air to pass from the oxygen chamber into the mouth, while air passing from the mouth back to the chamber is directed by means of another valve through the second tube Before the expired air from the mouth enters the upper portion of the oxygen chamber, it flows through a lower chamber containing pellets of soda lime, which combine chemically with the carbon dioxide in the expired air. Therefore, as oxygen is used by the person's body and the carbon dioxide is absorbed by the soda lime, the floating oxygen chamber, which is precisely balanced by a weight, gradually sinks in the water, owing to the oxygen loss. This chamber is coupled to a pen that records on a moving drum the rate at which the chamber sinks in the water and thereby records the rate at which the body utilizes oxygen.

FACTORS THAT AFFECT THE METABOLIC RATE

Factors that increase the chemical activity in the cells also increase the metabolic rate. Some of these are the following:

Exercise. The factor that causes by far the most dramatic effect on metabolic rate is strenuous exercise. Short bursts of maximal muscle contraction in any single muscle can liberate as much as a hundred times its normal resting amount of heat for a few seconds at a time. In considering the entire body, however, maximal muscle exercise can increase the overall heat production of the body for a few seconds to about 50 times normal or sustained for several minutes to about 20 times normal in the well-trained athlete, which is an increase in metabolic rate to 2000 per cent of normal.

Energy Requirements for Daily Activities. When

an average man of 70 kilograms lies in bed all day, he utilizes approximately 1650 Calories of energy. The process of eating increases the amount of energy utilized each day by an additional 200 or more Calories so that the same man lying in bed and also eating a reasonable diet requires a dietary intake of approximately 1850 Calories per day If he sits in a chair all day, his total energy requirement reaches 2000 to 2250 Calories. Therefore, in round figures, it can be assumed that the daily energy requirements simply for existing (that is, performing essential functions only) is about 2000 Calories.

Effects of Different Types of Work on Daily Energy Requirements. Table 71–1 illustrates the rates of energy utilization while one performs different types of activities Note that walking up stairs requires approximately 17 times as much energy as lying in bed asleep In general, over a 24-hour period a laborer can achieve a maximum rate of energy utilization as great as 6000 to 7000 Calories—in other words, as much as 3 5 times the basal rate of metabolism

Specific Dynamic Action of Protein. After a meal is ingested, the metabolic rate increases This is believed to result to a very slight extent from the different chemical reactions associated with digestion, absorption, and storage of food in the body However, it mainly results from the effect of directly stimulating the cellular chemical processes of certain of the amino acids derived from the proteins of the ingested food.

After a meal containing a large quantity of carbohydrates or fats, the metabolic rate usually increases only about 4 per cent However, after a meal containing large quantities of protein, the metabolic rate usually begins rising within 1 hour, reaches a maximum about 30 per cent above normal, and lasts for as long as 3 to 12 hours. This effect of protein on the metabolic rate is called the *specific dynamic action* of protein.

Age. The metabolic rate of a young child in relation to its size is almost two times that of an old person. This is illustrated later in Figure 71–5, which shows the declining metabolic rates of both males and females from birth until very old age. The high metabolic rate of young children results from high rates of cellular reactions, including partly the rapid synthesis of cellular materials and growth of the body, which require moderate quantities of energy.

Table 71–1 ENERGY EXPENDITURE PER HOUR *DURING DIFFERENT TYPES OF ACTIVITY* FOR A 70 KILOGRAM MAN

Form of Activity	Calories per Hour
Sleeping	65
Awake lying still	77
Sitting at rest	100
Standing relaxed	105
Dressing and undressing	118
Tailoring	135
Typewriting rapidly	140
"Light" exercise	170
Walking slowly (2 6 miles per hour)	200
Carpentry metal working, industrial painting	240
"Active" exercise	290
"Severe" exercise	450
Sawing wood	480
Swimming	500
Running (5.3 miles per hour)	570
"Very severe" exercise	600
Walking very fast (5.3 miles per hour)	650
Walking up stairs	1100

Extracted from data compiled by Professor M. S. Rose

Thyroid Hormone. When the thyroid gland secretes maximal quantities of thyroxine, the metabolic rate sometimes rises to as high as 100 per cent above normal. On the other hand, total loss of thyroid secretion decreases the metabolic rate to as low as 50 to 60 per cent of normal These effects can readily be explained by the basic function of thyroxine of increasing the rates of activity of almost all the chemical reactions in all cells of the body. This relationship between thyroxine and metabolic rate will be discussed in much greater detail in Chapter 76 in connection with thyroid function.

Sympathetic Stimulation—"Nonshivering Thermogenesis." Stimulation of the sympathetic nervous system, with liberation of norepinephrine and epinephrine, increases the metabolic rate of many tissues of the body. These hormones have a direct effect on muscle and liver cells in causing glycogenolysis, and this, along with other intracellular effects, increases cellular activity However, even more important is the effect of sympathetic stimulation on a certain type of fat tissue called *brown fat* in causing marked liberation of heat. This type of fat contains large numbers of mitochondria and many small globules of fat instead of one large fat globule In these cells, the process of oxidative phosphorylation in the mitochondria is mainly "uncoupled." That is, when the cells are stimulated by the sympathetic nerves their mitochondria produce a large amount of heat but almost no ATP, so that almost all the released oxidative energy becomes heat. The newborn child has a considerable number of such fat cells, and maximal sympathetic stimulation can increase the child's metabolism more than 100 per cent. This is called *nonshivering thermogenesis* The magnitude of this effect in the adult human being is in question—probably less than 10 to 15 per cent—though this might be increased following cold adaptation.

Male Sex Hormone. The male sex hormone can increase the basal metabolic rate about 10 to 15 per cent, and the female sex hormone perhaps a few per cent but usually not enough to be of great significance The difference in metabolic rates of males and females is illustrated in Figure 71–4

Growth Hormone. Growth hormone can increase the basal metabolic rate as much as 15 to 20 per cent as a result of direct stimulation of cellular metabolism.

Fever. Fever, regardless of its cause, increases the metabolic rate This is because all chemical reactions, either in the body or in the test tube, increase their rates of reaction an average of about 120 per cent for every 10° C rise in temperature. However, the body's temperature control system diminishes this effect somewhat, as we shall discuss in the following chapter.

Climate. Studies of metabolic rates of persons living in the different geographic zones have shown metabolic rates as much as 10 to 20 per cent lower in tropical regions than in arctic regions This difference is caused to some extent by adaptation of the thyroid gland, with increased secretion in cold climates and decreased secretion in hot climates. Indeed, far more persons develop hyperthyroidism in cold regions of the earth than in tropical regions.

Sleep. The metabolic rate falls 10 to 15 per cent below normal during sleep This fall is presumably due to two

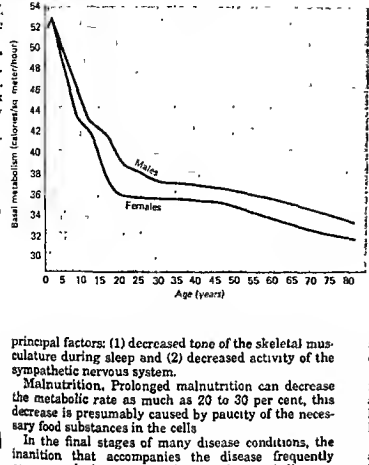


Figure 71–4. Normal basal metabolic rates at different ages for each sex.

principal factors: (1) decreased tone of the skeletal musculature during sleep and (2) decreased activity of the sympathetic nervous system.

Malnutrition. Prolonged malnutrition can decrease the metabolic rate as much as 20 to 30 per cent, this decrease is presumably caused by paucity of the necessary food substances in the cells

In the final stages of many disease conditions, the inanition that accompanies the disease frequently causes marked premortem decrease in metabolic rate, even to the extent that the body temperature may fall a number of degrees shortly before death.

THE BASAL METABOLIC RATE

The Basal Metabolic Rate as a Method for Comparing Metabolic Rates Between Individuals. It has been extremely important to establish a procedure that will measure the inherent activity of the tissues independently of exercise and other extraneous factors that would make it impossible to compare one person's metabolic rate with that of another person To do this, the metabolic rate is usually measured under so-called *basal conditions*, and the metabolic rate then measured is called the *basal metabolic rate* (BMR)

Basal Conditions. The basal metabolic rate means the rate of energy utilization in the body during absolute rest but while the person is awake. The following basal conditions are necessary for measuring the basal metabolic rate:

1. The person must not have eaten any food for at least 12 hours because of the specific dynamic action of foods.

2. The basal metabolic rate is determined after a night of restful sleep, for rest reduces the activities of the sympathetic nervous system and other metabolic excitants to their minimal level.

3. No strenuous exercise is performed after the night of restful sleep, and the person must remain at complete rest in a reclining position for at least 30 minutes prior to actual determination of the metabolic rate. This is perhaps the most important of all the conditions for attaining the basal state because of the extreme effect of exercise on metabolism.

4. All psychic and physical factors that cause excitement must be eliminated, and the subject must be made as comfortable as possible. These conditions, obviously, help to reduce the degree of sympathetic activity to as little as possible.

5. The temperature of the air must be comfortable and be somewhere between the limits of 68° and 80° F. Below 68° F, the sympathetic nervous system becomes progressively more activated to help maintain body heat, and above 80° F, discomfort, sweating, and other factors increase the metabolic rate.

Usual Technique for Determining the Basal Metabolic Rate. The usual method for determining basal metabolic rate is first to establish the subject under basal conditions and then to measure the rate of oxygen utilization using a metabolator of the type illustrated in Figure 71–3. Then the basal metabolic rate is calculated as shown in Figure 71–5

In the first few lines of Figure 71–5, the quantity of heat liberated in the body of the person is calculated from the quantity of oxygen utilized. In this figure, note that the total quantity of energy liberated in the body during the hour is 72.4 Calories.

Expressing the Basal Metabolic Rate in Terms of Surface Area. Obviously, if one subject is much larger than another, the total amount of energy utilized by the

15 liters — O_2 at standard conditions consumed in 1 hour
×4.825 — Calories liberated per liter of O_2 burned
72.4 — Calories liberated per hour
÷1.5 — Body surface area in square meters
48.3 — Calories per square meter per hour
−38.5 — Normal value for 20 year old man
9.8 — Excess Calories above normal

$$\frac{9.8 \times 100}{38.5} = 25.5 \text{ per cent above normal}$$

BMR = +25.5

Figure 71–5. Calculation of basal metabolic rate from the rate of oxygen consumption.

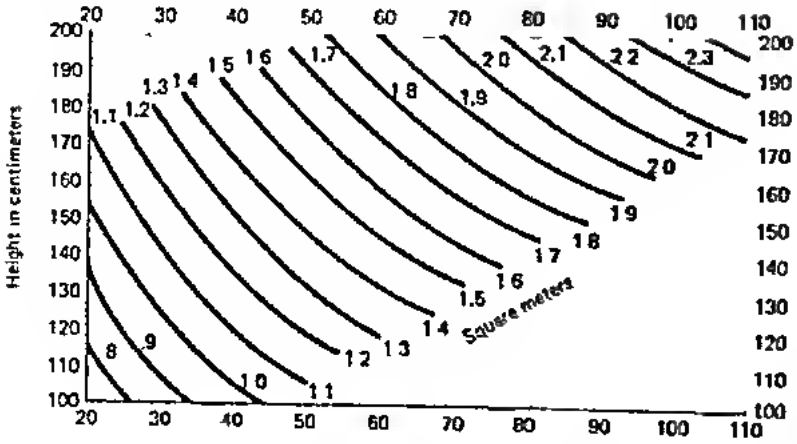


Figure 71–6. Relationship of height and weight to body surface area. (From DuBois Basal Metabolism in Health and Disease, Lea & Febiger)

two subjects will be considerably different simply because of differences in body size Experimentally, among normal persons, the average basal metabolic rate varies approximately *in proportion to the body surface area*

Note from Figure 71–5 that the total body surface area in this example is 1 5 square meters Therefore, the basal metabolic rate calculates to be 48 3 Calories per square meter per hour.

Method for Calculating the Total Surface Area. The surface area of the body varies approximately in proportion to *weight*$^{0.67}$ However, measurements of the body surface area have shown that it can be calculated more accurately by a complicated formula based on the weight and height of the subject as follows·

$$\text{Body surface area} = \text{Weight}^{0.425} \times \text{Height}^{0.725} \times 0.007184$$

Figure 71–6 presents a graph based on this formula when body surface area is expressed in *square meters*, weight in *kilograms*, and height in *centimeters*. The surface area of the average 70-kilogram adult is 1.73 square meters.

Expressing the Basal Metabolic Rate in Percentage Above or Below Normal. To compare the basal metabolic rate of any one subject with the normal basal metabolic rate, it is necessary to refer to a chart such as that in Figure 71–4, which gives the normal basal metabolic rate per square meter at each age for each sex. For example, if the person represented in the calculation of Figure 71–5 is a 20 year old man, then we find from Figure 71–4 that his normal metabolic rate is 38 5 Calories per square meter per hour. But his actual metabolic rate, 48 3 Calories, is 9 8 Calories per square meter per hour *above* the normal mean value It is then determined that this is 25 5 per cent above normal Therefore, the basal metabolic rate, usually called simply the "BMR," is expressed as *plus* 25 5 Similarly, basal metabolic rates below normal are expressed as minus values

Constancy of the Metabolic Rate in the Same Person. Basal metabolic rates have been measured in many subjects at repeated intervals for as long as 20 or more years As long as a subject remains healthy, almost invariably the BMR, when expressed as a percentage of

REFERENCES

Alexander, G Cold thermogenesis *In* Robertshaw D (ed) International Review of Physiology Environmental Physiology III Vol 20 Baltimore, University Park Press, 1979, p 43

Atkins, G L., An Outline of Energy Metabolism in Man Philadelphia, International Ideas, 1981

Baldwin, R L and Bywater A C Nutritional energetics of animals *Annu Rev Nutr* 4 101 1984

Becker D J The endocrine responses to protein-calorie malnutrition. *Annu Rev Nutr*, 3 187 1983

Bukowiecki, L J Mechanisms of stimulus calorigenesis coupling in brown adipose tissue *Can J Biochem Cell Biol*, 62 623, 1984

Durnin J V Energy balance in childhood and adolescence *Proc Nutr Soc*, 43 271, 1984

Golinick, P D Metabolism of substrates Energy substrate metabolism during exercise and as modified by training *Fed Proc*, 44 353, 1985.

Guyton, A C, and Farrish, C A A rapidly responding continuous oxygen consumption recorder *J Appl Physiol*, 14 143, 1959

Hornsher, E, and Kean C J Skeletal muscle energetics and metabolism *Annu Rev Physiol* 40 93, 1978

Magnen, J L Body energy balance and food intake A neuroendocrine regulatory mechanism *Physiol Rev*, 63 314 1983

Nicholls, D G, and Locke, R M · Thermogenic mechanisms in brown fat *Physiol Rev*, 64 1, 1984

Oppenheimer, J H · Thyroid hormone action at the nuclear level *Ann Intern Med*, 102 374, 1985

Storlein, L H The role of the ventromedial hypothalamic area in peripranDial glucoregulation *Life-Sci*, 36 505, 1985

Wyndham, C H, and Loots H Responses to cold during a year in Antarctica *J Appl Physiol*, 27 696, 1969

72

Body Temperature, Temperature Regulation, and Fever

"Core" Temperature and Surface Temperature. The temperature of the deep tissues of the body—the "core"—remains almost exactly constant, within ±1° F, day in and day out except when a person develops a febrile illness. Indeed, the nude person can be exposed to temperatures as low as 55° F or as high as 140° F in *dry* air and still maintain an almost constant internal body temperature. Obviously, the mechanisms for control of body temperatura represent a beautifully designed control system. The purpose of this chapter is to discuss this system as it operates in health and in disease.

The *surface temperature*, in contrast to the *core temperature*, rises and falls with tha temperature of the surroundings. This is the temperature that is important when we refer to the ability of the skin to lose heat to tha surroundings.

The Normal Body Temperature. No single temperature level can be considered to be normal, for measurements on many normal persons have shown a *range* of normal temperatures, as illustrated in Figure 72–1, from less than 97° F to over 99° F. When measured by rectum, the values are approximately 1° F greater than the oral temperatures. The average normal temperature is generally considered to be between 98.0° F and 98.6° F (36.7° C and 37° C) when measured orally and approximately 1° F or 0.6° C higher when measured rectally.

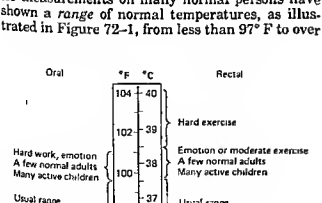


Figure 72–1. Estimated range of body temperature in normal persons. (from DuBois Fever. Charles C Thomas)

The body temperature varies somewhat with exercise and with extremes of temperature of tha surroundings, because the temperature regulatory mechanisms are not 100 per cent effectivs. When excessive heat is produced in tha body by strenuous exercise, the rectal temperature can risa to as bigh as 101° to 104° F. On the other hand, when tha body is exposed to cold, the rectal temperature can often fall to values considerably below 98° F.

THE INSULATOR SYSTEM OF THE BODY

The skin, the subcutaneous tissues, and the fat of the subcutaneous tissues are a heat insulator for the body. The fat is especially important because it conducts heat only *one third* as readily as other tissues. When no blood is flowing from the heated internal organs to the skin, the insulating properties of the normal male body are approximately equal to three quarters the insulating properties of a usual suit of clothes. In women this insulation is still better. Obviously, the degree of insulation varies, depending on the thickness of the subcutaneous fat; each millimeter of fat allows a person to be comfortable in a 2° to 3° F colder atmosphere.

Because most body heat is produced in the deeper portions of the body, the insulation beneath the skin is an effective means for maintaining normal internal core temperature, even though it allows the temperature of the skin to approach the temperature of the surroundings.

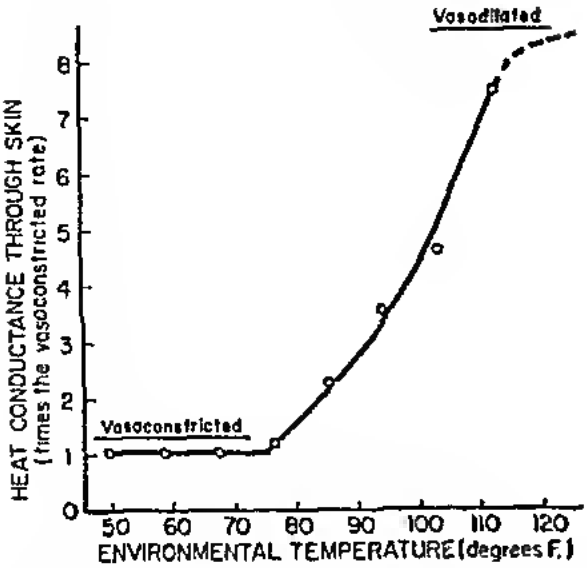


Figure 72–2. Effect of changes in the environmental temperature on the heat conductance from the body core to the skin surface (Modified from Benzinger Heat and Temperature, Fundamentals of Medical Physiology Dowden, Hutchinson and Ross, 1980)

FLOW OF BLOOD TO THE SKIN AND HEAT TRANSFER FROM THE BODY CORE

Blood vessels penetrate the subcutaneous insulator tissues and are distributed profusely in the subpapillary portions of the skin. Indeed, immediately beneath the skin is a continuous venous plexus that is supplied by inflow of blood from the skin capillaries And in the most exposed areas of the body—the hands, feet, and ears—blood is supplied to the plexus directly from the small arteries through highly muscular venous *artertovenous anastomoses* from the arterioles to the veins. The rate of blood flow into the venous plexus can vary tremendously—from barely above zero to as great as 30 per cent of the total cardiac output A high rate of blood flow causes heat to be conducted from the core of the body to the skin with great efficiency, whereas reduction in the rate of blood flow decreases the efficiency of heat conduction from the core Figure 72–2 shows quantitatively the effect of skin blood flow on conductance of heat from the body core to the skin surface, illustrating an opproximate eightfold increase in heat conductance between the fully vasoconstricted state ond the fully vasodilated state

Obviously, therefore, the skin is an effective "radiator" system, ond the flow of blood to the skin is a most effective mechanism of heat transfer from the body core to the skin

Control of Heat Conduction to the Skin by the Sympathetic Nervous System. Heat conduction to the skin by the blood is controlled by the degree of vasoconstriction of the arterioles and arteriovenous onastomoses that supply blood to the venous plexus of the skin. This vasoconstric- in turn is controlled almost entirely by the sympathetic nervous system in response t changes in the body core temperature and th environmental temperature. This will be discusse later in the chapter in connection with the contro of body temperature by the hypothalamus.

BALANCE BETWEEN HEAT PRODUCTION AND HEAT LOSS

Heat is continually being produced in the bod as a by-product of metabolism, and body heat also continually being lost to the surrounding When the rate of heat production is exactly equ to the rate of loss, the person is said to be in he balance But when the two are out of equilibriu the body heat, and the body temperature as wel will obviously be either increasing or decreasing

The important factors that play major roles i determining the rate of *heat production* were di cussed in detail in Chapter 71; the most importa of these may be listed as follows: (1) basal rate metabolism of all the cells of the body; (2) increa in rate of metabolism caused by muscle activit including that caused by shivering; (3) increase i metabolism caused by the effect of thyroxine c cells, (4) increase in metabolism caused by t effect of epinephrine, norepinephrine, and symp thetic stimulation on cells; and (5) increase metabolism caused by increased temperature the body cells.

HEAT LOSS

The various methods by which heat is lost from the body are pictured in Figure 72–3. These include *radiation, conduction,* and *evaporation* and may be explained as follows:

Radiation. As illustrated in Figure 72–3, a nude person in a room at normal room temperature loses about 60 per cent of the total heat loss by radiation

Loss of heat by radiation means loss in the form of infrared heat rays, a type of electromagnetic wave Most infrared heat rays radiating from the body have wavelengths of 5 to 20 microns, 10 to 30 times the wavelengths of light rays. All objects that are not at absolute zero temperature radiate such rays. Therefore, the human body radiates

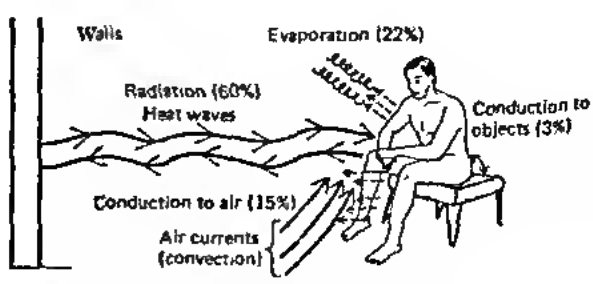


Figure 72–3. Mechanisms of heat loss from the body

heat rays in all directions. However, heat rays are also being radiated from the walls and other objects toward the body. If the temperature of the body is greater than the temperature of the surroundings, a greater quantity of heat is radiated from the body than is radiated to the body.

Conduction. Usually, as illustrated in Figure 72–3, only minute quantities of heat are lost from the body by direct conduction from the surface of the body *to other objects,* such as a chair or a bed. When one first sits on a chair while nude, heat is conducted from the body to the chair rapidly, but within a few minutes the temperature of the chair rises almost to equal the temperature of the body, and thereafter the chair actually becomes an insulator to prevent much additional loss of heat.

On the other hand, loss of heat by *conduction to air* does represent a sizeable proportion of the body's heat loss even under normal conditions It will be recalled that heat is actually the kinetic energy of molecular motion, and the molecules that comprise the skin of the body are continually undergoing vibratory motion. The energy of this motion can be transferred to the air if the air is colder than the skin, thus increasing the velocity of motion of the air molecules. However, once the temperature of the air immediately adjacent to the skin equals tha temperatura of the skin, no further loss of heat from tha body to the air can occur Therefore, conduction of heat from the body to the air is self-limited *unless the heated air moves away from the skin* so that new, unheated air is continually brought in contact with the skin, a phenomenon called air convection.

Convection. The removal of heat from the body by convection air currents is commonly called "heat loss by convection." Actually, the heat must first be *conducted* to the air and then carried away by the convection currents.

A small amount of convection almost always occurs around the body because of the tendency for the air adjacent to the skin to rise as it becomes heated. Therefore, a nude person seated in a comfortable room without gross air movement still loses about 12 per cent of his or her heat by conduction to the air and then by air convection away from the body.

Cooling Effect of Wind. When the body is exposed to wind, the layer of air immediately adjacent to the skin is replaced by new air much more rapidly than normally, and heat loss by convection increases accordingly. The cooling effect of wind at low velocities is approximately proportional to the square root of the wind velocity. For instance, a wind of 4 miles per hour is about two times as effective for cooling as a wind of 1 mile per hour.

Conduction and Convection of Heat from a Person Exposed to Water. Water has a specific heat several thousand times as great as that of air so that each unit portion of water adjacent to the skin can absorb far greater quantities of heat than can air. Also, the conductivity of water for heat is marked in comparison with that of air. Consequently, it is impossible to heat a thin layer of water next to the body to form an "insulator zone" as occurs in air. Therefore, the rate of heat loss to water at moderate temperatures is many times as great as the rate of heat loss to air of the same temperature. However, when both the water and air are extremely cold, the rate of heat loss to air becomes almost as great as to water, for both water and air are then capable of carrying away essentially all the heat that can diffuse through the subcutaneous insulation of the skin.

Evaporation. When water evaporates from the body surface, 0.58 Calorie (kilocalorie) of heat is lost for each gram of water that evaporates. Water evaporates *insensibly* from the skin and lungs at a rate of about 600 ml per day. This causes continual heat loss at a rate of 12 to 16 Calories per hour. Unfortunately, this insensible evaporation of water directly through the skin and lungs cannot be controlled for purposes of teroperature regulation because it results from continual diffusion of water molecules through the skin and respiratory surfaces regardless of body temperature. However, loss of heat by evaporation of sweat on the skin surface can be controlled by regulating the rate of sweating, which is discussed later in tha chapter.

Evaporation as a Necessary Refrigeration Mechanism at Very High Air Temperatures. As long as skin temperature is greater than the temperature of the surroundings, heat is lost by radiation and conduction, but when the teroperature of the surroundings is greater than that of the skin, instead of losing heat the body gains heat by radiation and conduction. Under these conditions, *the only means by which the body can rid itself of heat is by evaporation.* Therefore, any factor that prevents adequate evaporation when the surrounding temperatures are higher than skin temperature will cause the body tsmperature to rise. This occurs occasionally in human beings who are born with congenital absence of sweat glands. These persons can stand cold temperatures as well as can normal persons, but they are likely to die of heat stroke in tropical zones, for without the evaporative refrigeration system they cannot prevent a rise in body temperature when the air temperature is above that of the body.

Effect of Clothing on Heat Loss. ***Effect on Conductive Heat Loss.*** Clothing entraps air next to the skin and in the weave of the cloth, thereby increasing the thickness of the so-called "private zone" of air adjacent to the skin and decreasing the flow of convection air currents. Consequently, the rate of heat loss from the body by conduction and convection is greatly depressed. A usual suit of clothes decreases the rate of heat loss to about half that from a nude body, whereas arctic-type clothing can decrease this heat loss to as little as one sixth.

About half the heat transmitted from the skin to the clothing is radiated to the clothing instead of being conducted across the small intervening space. Therefore, coating the inside of clothing with a thin layer of gold, which reflects radiant heat back to the body, makes the insulating properties of clothing far more effective than

otherwise. Using this technique, clothing for use in the arctic can be decreased in weight by about half.

The effectiveness of clothing in preventing heat loss is almost completely lost when it becomes wet, because the high conductivity of water increases the rate of heat transmission as much as 20-fold or more. One of the most important factors for protecting the body against cold in arctic regions is extreme caution against allowing the clothing to become wet. Indeed, one must be careful not to overheat oneself even temporarily, for sweating in one's clothes makes them much less effective thereafter as an insulator.

SWEATING AND ITS REGULATION BY THE AUTONOMIC NERVOUS SYSTEM

Stimulation of the preoptic area in the anterior part of the hypothalamus either electrically or by excess heat causes sweating. The impulses from this area that cause sweating are transmitted in the autonomic pathways to the cord and thence through the sympathetic outflow to the skin everywhere in the body.

It should be recalled from the discussion of the autonomic nervous system in Chapter 57 that the sweat glands are innervated by sympathetic *cholinergic* nerve fibers. However, these glands can also be stimulated by epinephrine or norepinephrine circulating in the blood even though the glands themselves, in most parts of the body, do not have adrenergic innervation. It is possible that the sweat glands of the hands and feet do have some adrenergic innervation as well as cholinergic innervation, for many emotional states that excite the adrenergic portions of the sympathetic nervous system are known also to cause local sweating in these areas. Also, during exercise, which normally excites adrenergic activity, localized sweating of the hands and feet occurs. In this instance, the moisture from the sweat helps the surfaces of the hands and feet to gain traction against smooth surfaces and also prevents drying of the thick cornified layers of skin.

Mechanism of Sweat Secretion. Figure 72–4 illustrates a sweat gland, showing it to be a tubular structure consisting of two parts: (1) a deep subdermal *coiled portion* that secretes the sweat, and (2) a *duct portion* passing outward through the dermis and epidermis of the skin. As is true of so many other glands, the secretory portion of the sweat gland secretes a fluid called the *primary secretion* or *precursor secretion;* then the concentrations of the constituents in the fluid are modified as the fluid flows through the duct.

The precursor secretion is an active secretory product of the epithelial cells lining the coiled portion of the sweat gland. Cholinergic sympathetic nerve fibers ending on or near the glandular cells elicit the secretion.

The composition of the precursor secretion is very similar to that of plasma except that it does not contain the plasma proteins. The concentration of sodium is about 142 mEq per liter and chloride

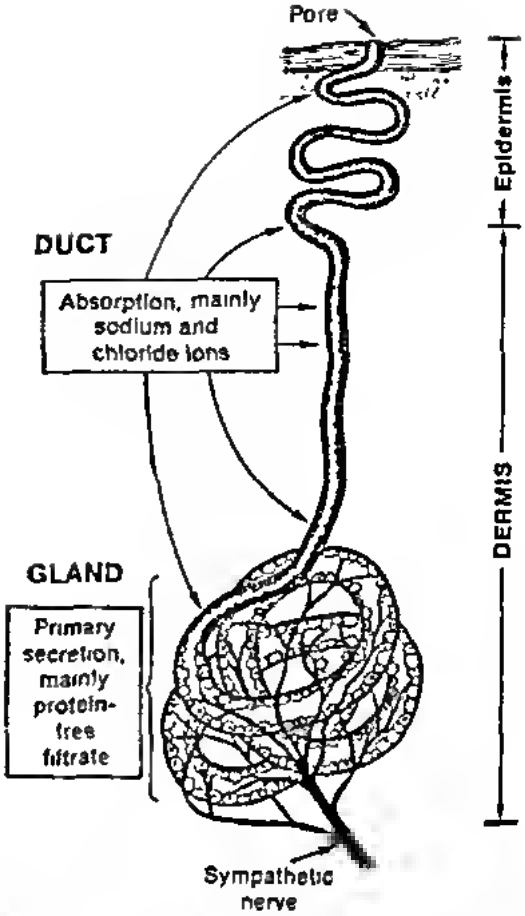


Figure 72–4. A sweat gland innervated by a sympathetic nerve. A *primary secretion* is formed by the glandular portion, but much if not most of the electrolytes are reabsorbed in the duct, leaving a dilute, watery secretion.

about 104 mEq per liter, with much smaller concentrations of the other solutes of plasma. However, as this precusor solution flows through the duct portion of the gland, it is modified by reabsorption of most of the sodium and chloride ions. The degree of this reabsorption depends upon the rate of sweating as follows:

When the sweat glands are stimulated only slightly, the precursor fluid passes through the duct very slowly. In this instance, essentially all the sodium and chloride ions are reabsorbed, and the concentration of each of these falls to as low as 5 mEq per liter. This reduces the osmotic pressure of the fluid to such a low level that most of the water is also reabsorbed, which concentrates most of the other constituents. Therefore, at low rates of sweating, such constituents as urea, lactic acid, and potassium ions are usually very concentrated.

On the other hand, when the sweat glands are strongly stimulated by the sympathetic nervous system, very large amounts of precursor secretion are formed, and the duct can now reabsorb only slightly more than one half of the sodium chloride; the concentrations of the sodium and chloride ions then usually rise (in the *unacclimatized* person) to

maximum levels of about 60 mEq per liter, slightly less than half the concentrations in plasma. Furthermore, the water reabsorption is greatly reduced so that the other constituents of the sweat are only moderately increased in concentration: urea is then about two times that in the plasma, lactic acid about four times, and potassium about 1.2 times.

Note especially the large loss of sodium chloride in the sweat when the person is unacclimatized. However, this is quite different once the person has become acclimatized to heat as follows:

Acclimatization of the Sweating Mechanism. Though a normal, unacclimatized person can rarely produce more than about 700 ml of sweat per hour, when exposed to hot weather for one to six weeks the person sweats progressively more profusely, often increasing the maximum sweat production to as much as 1.5 to 2 liters per hour Evaporation of this much sweat can remove heat from the body at a rate *more than ten times* the normal basal rate of heat production. This increased effectiveness of the sweating mechanism is caused by a direct increase in sweating capability of the sweat glands themselves.

Also associated with the acclimatization is decreased concentration of sodium chloride in the sweat, which allows progressively better conservation of salt. Most of this effect is caused by *increased secretion of aldosterone*, which, in turn, results from decreasing sodium chloride in the body. An unacclimatized person who sweats profusely often loses as much as 15 to 30 grams of salt each day for the first few days But, after four to six weeks of acclimatization, the loss will usually be as little as 3 to 5 grams per day

Panting as a Means of Evaporative Heat Loss. Many lower animals do not have sweat glands, but to offset this they lose large amounts of heat by the panting mechanism. During panting, only small volumes of air pass in and out of the lungs with each breath so that mainly dead space air enters the alveoli. Therefore, the animal does not become overventilated, but nevertheless tremendous amounts of air are moved over the surfaces of the tongue, mouth, and trachea. Evaporation from these respiratory surfaces, especially of saliva on the tongue, is an important mechanism for body heat control.

A special nervous center in the pons controls panting This center modifies the normal respiratory pattern to provide the rapid and shallow breathing required for the panting mechanism.

REGULATION OF BODY TEMPERATURE—THE "HYPOTHALAMIC THERMOSTAT"

Figure 72–5 illustrates approximately what happens to the temperature of the nude body after a few hours' exposure to *dry* air ranging from 30° to 170° F. Obviously, the precise dimensions of this curve vary, depending on the movement of air, the

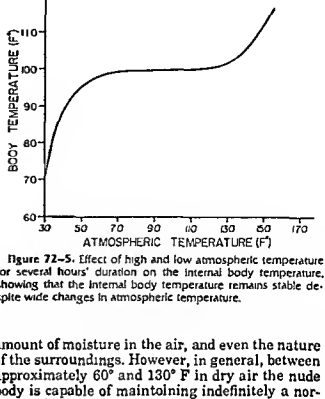


Figure 72–5. Effect of high and low atmospheric temperature for several hours' duration on the internal body temperature, showing that the internal body temperature remains stable despite wide changes in atmospheric temperature.

amount of moisture in the air, and even the nature of the surroundings. However, in general, between approximately 60° and 130° F in dry air the nude body is capable of maintaining indefinitely a normal body core temperature somewhere between 97° and 100° F.

The temperature of the body is regulated almost entirely by nervous feedback mechanisms, and almost all of these operate through *temperature-regulating centers* located in the *hypothalamus*. However, for these feedback mechanisms to operate, there must also exist temperature detectors, discussed next, to determine when the body temperature becomes either too hot or too cold.

Thermostatic Detection of Temperature in the Hypothalamus—Role of the Preoptic Area

In recent years, experiments have been performed in which minute areas in the brain have been either heated or cooled by use of a so-called *thermode*. This small, needle-like device is heated by electrical means or by passing hot water through it, or it is cooled by cold water. The principal area in the brain in which heat from a thermode affects body temperature control is the preoptic area of the hypothalamus, and to a lesser extent the adjacent regions of the anterior hypothalamus.

Using the thermode, the preoptic area of the hypothalamus has been found to contain large numbers of heat-sensitive neurons that seem to function as temperature sensors for controlling body temperature. These neurons increase their firing rate as the temperature rises, the rate sometimes increasing as much as tenfold with an increase in body temperature of 10° C.

In addition to the heat-sensitive neurons, a few

cold-sensitive neurons have also been found in the hypothalamus, in the septum, and in the reticular substance of the midbrain, all of which increase their rate of firing when exposed to cold. Also, still other neurons change their rates of firing in response to signals transmitted to the brain from cold and warmth receptors in the skin and certain deep tissues of the body. These signals are believed to interact with those from both the heat- and cold-sensitive neurons of the hypothalamus to provide the very effective temperature control system of the body

When the preoptic area is heated, the skin immediately breaks out into a profuse sweat while at the same time the skin blood vessels over the entire body become greatly vasodilated Thus, this is an immediate reaction to cause the body to lose heat, thereby helping to return the body temperature toward the normal level In addition, excess body heat production is inhibited. Therefore, it is clear that the preoptic area of the hypothalamus has the capability of serving as a thermostatic body temperature control center

Detection of Temperature by Receptors in the Skin and Deep Body Tissues

Though the signals generated by the temperature receptors of the hypothalamus are extremely powerful in controlling body temperature, temperature receptors in other parts of the body also play important roles in temperature regulation. This is especially true of temperature receptors in the skin and in a few specific deep tissues of the body.

It will be recalled from the discussion of sensory receptors in Chapter 49 that the skin is endowed with both *cold* and *warmth* receptors However, there are far more cold receptors than warmth receptors; in fact, ten times as many in many parts of the skin Therefore, peripheral detection of temperature mainly concerns detecting cool and cold instead of warm temperatures

When the skin is chilled over the entire body, immediate reflex effects are invoked to increase the temperature of the body in several ways: (1) by providing a strong stimulus to cause shivering, with resultant increase in the rate of body heat production, (2) by inhibiting the process of sweating if this should be occurring, and (3) by promoting skin vasoconstriction to diminish the transfer of body heat to the skin.

The deep body temperature receptors are found only in certain parts of the body, mainly in the *spinal cord*, in the *abdominal viscera*, and in or around the *great veins*. However, these deep receptors function differently from the skin receptors, for they are exposed to the body core temperature rather than the body surface temperature. Yet, like the skin temperature receptors, they mainly detect cold rother than warmth. It is probable that both the skin and the deep body receptor concerned with preventing hypothermia—th preventing low body temperatures.

Integration of the Temperature Signals in the Posterior Hypothalamus

Even though a large share of the signals for temperature detection arise in peripheral receptors, these signals help control body temperature mainly through the hypothalamus. However, the area of the hypothalamus that they stimulate is not the preoptic area but instead an area located bilaterally in the posterior hypothalamus approximately at the level of the mammary bodies. The thermostatic signals from the preoptic area are also transmitted into this posterior hypothalamus area Here the signals from the preoptic area and the signals from the body periphery are combined to provide the heat-producing or heat-losing reactions of the body.

The overall heat-controlling mechanism of the hypothalamus is called the *hypothalamic thermostat.*

NEURONAL EFFECTOR MECHANISMS THAT DECREASE OR INCREASE BODY TEMPERATURE

When the hypothalamic thermostat detects that the body temperature is either too hot or too cold, it institutes appropriate temperature-decreasing or temperature-increasing procedures. The student is familiar with most of these from personal experience, but special features are the following:

Temperature-Decreasing Mechanisms

The thermostatic system employs three important mechanisms to reduce body heat when the temperature becomes too great:

1 *Vasodilatation.* In almost all areas of the body the skin blood vessels are intensely dilated. This is caused by *inhibition of the sympathetic centers in the posterior hypothalamus that cause vasoconstriction* Full vasodilatation can increase the rate of heat transfer to the skin as much as eightfold.

2 *Sweating.* The effect of increased temperature on causing sweating is illustrated by the solid curve in Figure 72–6, which shows a sharp increase in the rate of evaporative heat loss resulting from sweating when the body core temperature rises above the critical temperature level of 37° C (98 6° F). An additional 1° C increase in body temperature causes enough sweating to remove ten times the basal rate of body heat production.

3. *Decrease in heat production.* Those mechanisms that cause excess heat production, such as shivering and chemical thermogenesis, are strongly inhibited.

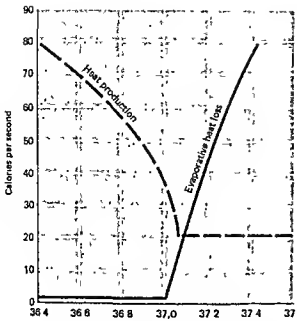


Figure 72–6. Effect of hypothalamic temperature on (1) evaporative heat loss from the body and (2) heat production caused primarily by muscular activity and shivering. This figure demonstrates the extremely critical temperature level at which increased heat loss begins and increased heat production stops. (Drawn from data in Benzinger, Kitzinger, and Pratt, in Hardy [ed.] Temperature, Part 3, p. 637, Reinhold Publishing Corp.)

Temperature-Increasing Mechanisms

When the body is too cold, the temperature control system institutes exactly opposite procedures. These are:

1. *Skin vasoconstriction throughout the body.* This is caused by stimulation of the posterior hypothalamic sympathetic centers.

2. *Piloerection.* Piloerection means hairs "standing on end." Sympathetic stimulation causes the arrector pili muscles attached to the hair follicles to contract, which brings the hairs to an upright stance. This is not important in the human being, but in lower animals upright projection of the hairs allows them to entrap a thick layer of "insulator air" next to the skin so that the transfer of heat to the surroundings is greatly depressed.

3. *Increase in heat production.* Heat production by the metabolic systems is increased by promoting (a) shivering, (b) sympathetic excitation of heat production, and (c) thyroxine secretion. These require additional explanation, as follows:

Hypothalamic Stimulation of Shivering. Located in the dorsomedial portion of the posterior hypothalamus near the wall of the third ventricle is an area called the *primary motor center for shivering*. This area is normally inhibited by signals from the heat center in the preoptic thermostatic area but is excited by cold signals from the skin and spinal cord. Therefore, as illustrated by the dashed curve in Figure 72–6, this center becomes activated when the body temperature falls even a fraction of a degree below a critical temperature level. It then transmits signals that cause shivering through bilateral tracts down the brain stem, into the lateral columns of the spinal cord, and, finally, to the anterior motor neurons. These signals are nonrhythmic and do not cause the actual muscle shaking. Instead, they increase the tone of the skeletal muscles throughout the body. When the tone rises above a certain critical level, shivering begins. This probably results from feedback oscillation of the muscle spindle stretch reflex mechanism, which was discussed in Chapter 51. During maximum shivering, body heat production can rise to as high as four to five times normal.

Sympathetic "Chemical" Excitation of Heat Production. It was pointed out in Chapter 71 that either sympathetic stimulation or circulating norepinephrine and epinephrine in the blood can cause an immediate increase in the rate of cellular metabolism; this effect is called *chemical thermogenesis*, and it results at least partially from the ability of norepinephrine and epinephrine to uncouple oxidative phosphorylation, as a result of which more oxidation of foodstuffs must occur to produce the energy required for normal function of the body. Therefore, the rate of cellular metabolism increases.

The degree of chemical thermogenesis that occurs in an animal is almost directly proportional to the amount of *brown* fat that exists in the animal's tissues. This is a type of fat that contains large numbers of mitochondria in its cells, as described in Chapter 71, and these cells are supplied by a strong sympathetic innervation.

The process of acclimatization greatly affects the intensity of chemical thermogenesis; some animals that have been exposed for several weeks to a very cold environment exhibit as much as a 100 to 500 per cent increase in heat production when acutely exposed to cold, in contrast to the unacclimatized animal, which responds with an increase of perhaps one third as much.

In adult human beings, who have almost no brown fat, it is rare that chemical thermogenesis increases the rate of heat production more than 10 to 15 per cent. However, in infants, who *do* have a small amount of brown fat in the interscapular space, chemical thermogenesis can increase the rate of heat production as much as 100 per cent, which is probably a very important factor in maintaining normal body temperature in the neonate.

Increased Thyroxine Output as a Cause of Increased Heat Production. Cooling the preoptic area of the hypothalamus also increases the production of the neurosecretory hormone *thyrotropin-releasing hormone* by the hypothalamus. This hormone is carried by way of the hypothalamic portal veins to the anterior pituitary gland where it stimulates the secretion of thyroid-stimulating hormone. Thyroid-stimulating hormone, in turn, stimulates increased output of thyroxine by the

thyroid gland, as will be explained in Chapter 76. The increased thyroxine increases the rate of cellular metabolism throughout the body, which is yet another mechanism of *chemical thermogenesis*. However, this increase in metabolism does not occur immediately but requires several weeks for the thyroid gland to hypertrophy before it reaches its new level of thyroxine secretion.

Exposure of animals to extreme cold for several weeks can cause their thyroid glands to increase in size as much as 20 to 40 per cent. However, human beings rarely allow themselves to be exposed to the same degree of cold as that to which animals have been subjected. Therefore, we still do not know, quantitatively, how important the thyroid method of adaptotion to cold is in the human being Yet, isolated measurements have shown that military personnel residing for several months in the Arctic develop increased metabolic rates; Eskimos also have abnormally high basal metabolic rates. Also, the continuous stimulatory effect of cold on the thyroid gland can probably explain the much higher incidence of toxic thyroid goiters in persons living in colder climates than in those living in warmer climates.

THE CONCEPT OF A "SET-POINT" FOR TEMPERATURE CONTROL

In the example of Figure 72–6, it is clear that at a very critical body core temperature, at a level of almost exactly 37.1° C, drastic changes occur in both the rate of heat loss and heat production. At temperatures above this level, the rate of heat loss is greater than heat production, so that the body temperature falls and reapproaches the 37.1° C level At temperotures below this level, the rate of heat production is greater than heat loss, so that now the body temperature rises and again approaches the 37.1° C level. Therefore, this critical *temperature level is called the "set-point"* of the temperature control mechanism. That is, all the temperature control mechanisms continually attempt to bring the body temperoture back to this "set-point" level

The Feedback Gain for Body Temperature Control. Let us recall for a moment the discussion of feedback gain of control systems that was presented in Chapter 1. Feedback gain is a measure of the effectiveness of a control system. In the case of body temperature control, it is important for the internal body temperature to change as little as possible despite marked changes in the environmental temperature, and the gain of the temperature control system is approximately equal to the ratio of the change in environmental temperature to the change in body temperoture that this causes. Experiments have shown that the body temperature of man changes about 1° C for each 25° to 30° C change in environmental temperature. Therefore, the feedback gain of the hypothalamic ther- mechanism for control of body temperature averages about 27, which is an extremely high gain for a biological control system (the baroreceptor arterial pressure control system, for instance, has a gain of less than 2).

Effect of Skin and Deep Body Temperatures in Altering the Temperature Level of the Set-Point

The critical temperature set-point in the hypothalamus above which sweating begins and below which shivering begins is determined mainly by the degree of activity of the heat temperature receptors in the preoptic area of the hypothalamus. However, we have already pointed out that temperature signals from the peripheral areas of the body, especially from the skin and certain deep body tissues (the spinal cord and the abdominal viscera) also contribute very significantly to body temperature regulation. But, how do they contribute? The answer is that they alter the set-point of the hypothalamic thermostat. This effect is illustrated in Figures 72–7 and 72–8.

Figure 72–7 illustrates the effect of different skin temperatures on the set-point, showing this to increase as the skin temperature decreased Thus, in the person represented in this figure, the hypothalamic set-point increased from 36.7° C when the skin temperature was higher than 33° C to a set-point of 37.4° C when the skin temperature had fallen ta 29° C. Therefore, when the skin temperature was very high, sweating began at a much lower hypothalamic temperature than when the skin temperature was low. One can readily

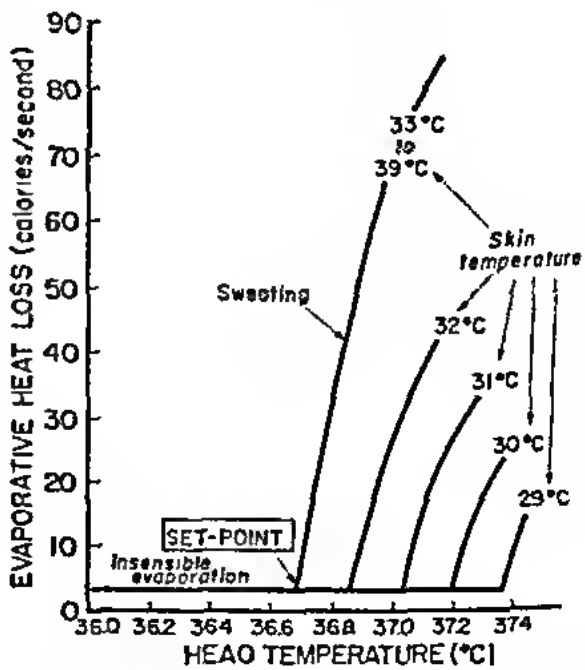


Figure 72–7. Effect of changes in the internal head temperature on the rate of evaporative heat loss from the body Note also that the skin temperature determines the exact "set-point" level at which sweating begins. (Courtesy of Dr T H. Benzinger)

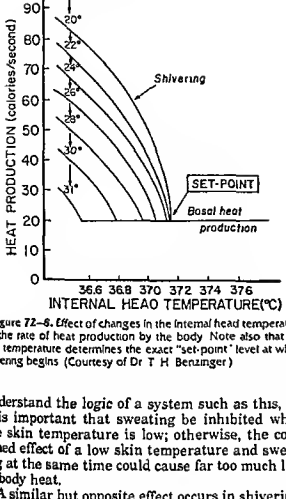


Figure 72–8. Effect of changes in the internal head temperature on the rate of heat production by the body. Note also that the skin temperature determines the exact "set-point" level at which shivering begins. (Courtesy of Dr. T. H. Benzinger.)

understand the logic of a system such as this, for it is important that sweating be inhibited when the skin temperature is low; otherwise, the combined effect of a low skin temperature and sweating at the same time could cause far too much loss of body heat.

A similar but opposite effect occurs in shivering. Note in Figure 72–8 that shivering begins at a hypothalamic temperature of approximately 37.1° C when the skin temperature is a very cold 20° C; but when the skin is at a warm level of 31° C, the hypothalamic temperature must fall almost to 36.5° C before shivering begins. That is, when the skin becomes cold it drives the hypothalamic thermostat to the shivering threshold even when the hypothalamic temperature itself is still quite hot. Here again, one can well understand the logic of the control system, because a cold skin temperature will soon lead to deeply depressed body temperature unless heat production is increased. Thus, this effect of a cold skin temperature on increasing heat production actually "anticipates" a possible fall in internal body temperature and prevents its occurrence.

BEHAVIORAL CONTROL OF BODY TEMPERATURE

Aside from the thermostatic mechanism for body temperature control, the body has still another temperature-controlling mechanism that is even more potent than all the previously cited mechanisms together. This is behavioral control of temperature, which can be explained as follows: Whenever the internal body temperature becomes too high, signals from the thermostatic brain areas give the person a psychic sensation of being overheated. Conversely, whenever the body becomes too cold, signals from the skin and probably also from the deep body receptors elicit the feeling of cold discomfort. Therefore, the person makes appropriate environmental adjustments to re-establish comfort. This is a much more powerful system of body temperature control than most physiologists have recognized in the past. Indeed, for people, this is the only really effective mechanism for body heat control in severely cold environs.

The obvious types of behavioral adjustments include selecting appropriate clothing, moving the body to a different environmental setting, increasing the delivery of heat or cold from appropriate heaters or air conditioners, and so forth.

It is important to note that many other of our body's control systems utilize similar behavior mechanisms to achieve highly refined degrees of control. For instance, even respiration is controlled to a great extent in this way—that is, when people perceive that they are being subjected to air hunger, they consciously breathe more to make up the deficit. Therefore, it is not valid to think of the body's homeostatic control systems as operating only in the subconscious portions of the brain.

LOCAL SKIN TEMPERATURE REFLEXES

When a person places a foot under a hot lamp and leaves it there for a short time, *local vasodilatation* and mild *local sweating* occur. Conversely, placing the foot in cold water causes local vasoconstriction and local cessation of sweating. These reactions are caused by local effects of temperature directly on the blood vessels and sweat glands and also by local cord reflexes conducted from the skin receptors to the spinal cord and back to the same skin area. However, the *intensity* of these local effects is controlled by the hypothalamic thermostat, so that the overall effect is approximately proportional to the hypothalamic heat control signal *times* the local signal. Such reflexes can help prevent excessive heat exchange from locally cooled or heated portions of the body.

Regulation of Internal Body Temperature After Cutting the Spinal Cord. After cutting the spinal cord in the neck above the sympathetic outflow from the cord, regulation of body temperature becomes extremely poor, for the hypothalamus can then no longer control either skin blood flow or the degree of sweating *anywhere* in the body. On the other hand, the local temperature reflexes originating in the skin, spinal cord, and intra-abdominal receptors still exist. Unfortunately, these reflexes are not powerful. In persons with this condition, body temperature must be regulated principally by the patients' psychic response to cold and hot sensations in the head region—that is, by behavioral control. If they feel themselves becoming too hot or if they develop a headache from the heat, they know to select cooler surroundings, and, conversely, if they have cold sensations, they select warmer surroundings.

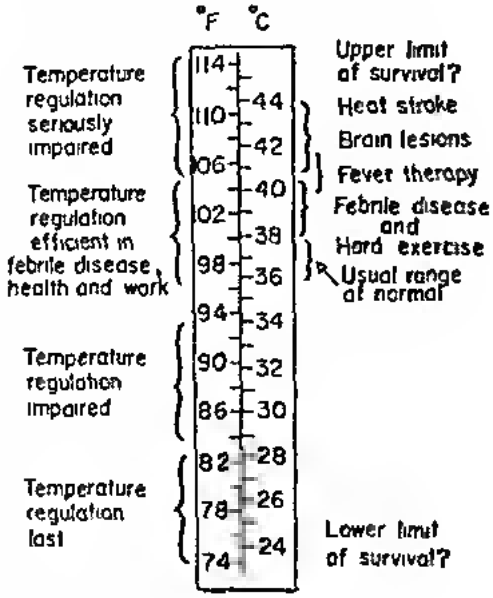


Figure 72–9. Body temperatures under different conditions (From DuBois Fever Charles C Thomas)

ABNORMALITIES OF BODY TEMPERATURE REGULATION

FEVER

Fever, which means a body temperature above the usual range of normal, may be caused by abnormalities in the brain itself or by toxic substances that affect the temperature-regulating centers. Some causes of fever are presented in Figure 72–9 These include bacterial diseases, brain tumors, and environmental conditions that may terminate in heat stroke

Resetting the Hypothalamic Thermostat in Febrile Diseases—Effect of Pyrogens

Many proteins, breakdown products of proteins, and certain other substances, especially lipopolysaccharide toxins secreted by bacteria, can cause the set-point of the hypothalamic thermostat to rise Substances that cause this effect are called *pyrogens*. It is pyrogens secreted by toxic bacteria or pyrogens released from degenerating tissues of the body that cause fever during disease conditions When the set-point of the hypothalamic thermostat becomes increased to a higher level than normal, all the mechanisms for raising the body temperature are brought into play, including heat conservation and increased heat production. Within a few hours after the thermostat has been set to a higher level, the body temperature also approaches this level.

Mechanism of Action of Pyrogens in Causing Fever—Role of "Endogenous Pyrogen." Experiments in animals have shown that most pyrogens, when injected into the hypothalamus, can act directly on the hypothalamic thermostat to increase its set-point, though many pyrogens require several hours of latent period before causing this effect. However, many of the bacterial pyrogens, especially the *endotoxins* from gram-negative bacteria, can also cause very severe fever acting in an indirect manner rather than directly on the hypothalamus, as follows:

When bacteria or breakdown products of bacteria are present in the tissues or in the blood, these are *phagocytized by the blood leukocytes and the tissue macrophages*. Both these types of cells in turn digest the bacterial products and then release into the body fluids a substance called *leukocyte pyrogen* or *endogenous pyrogen*. This endogenous pyrogen, on reaching the hypothalamus, immediately produces fever, increasing the body temperature in as little as 8 to 10 minutes. As little as one ten millionth of a gram of endotoxin lipopolysaccharide acting in this manner in concert with the blood leukocytes and the tissue macrophages can cause fever. The amount of endogenous pyrogen that is formed in response to the lipopolysaccharide to cause the fever is only a few nanograms

Several recent experiments have suggested that endogenous pyrogen causes fever by first inducing the formation of *prostaglandin* E_1 in the local cells of the hypothalamus; the prostaglandin E_1 in turn elicits the fever reaction When prostaglandin formation is blocked by drugs, the fever is either completely abrogated or at least reduced. In fact this may be the explanation for the manner in which aspirin reduces the degree of fever, because aspirin impedes the formation of prostaglandins It also would explain why aspirin does not lower the body temperature in a normal person, because a normal person does not have any endogenous pyrogen affecting the hypothalamus. Drugs such as aspirin that redure the level of fever are called *antipyretics*.

Fever Caused by Brain Lesions. When a brain surgeon operates in the region of the hypothalamus, this almost invariably causes severe fever though rarely the opposite effect occurs, thus illustrating both the potency of the hypothalamic thermostat mechanism for body temperature control and also the ease with which abnormalities of the hypothalamus can alter the set-point of the thermostat. Another condition that frequently causes prolonged high temperature is compression of the hypothalamus by brain tumors.

Characteristics of Febrile Conditions

Chills. When the set-point of the hypothalamic thermostat is suddenly changed from the normal level to a higher than normal value as a result of tissue destruction, pyrogenic substances, or dehydration, the body temperature usually takes several hours to reach the new temperature set-point. Figure 72–10 illustrates this, showing the effect of suddenly increasing the set point to a level of 103° F. Because the blood temperature is less than the set-point of the hypothalamic thermostat, the usual responses that cause elevation of body temperature occur. During this period the person exp

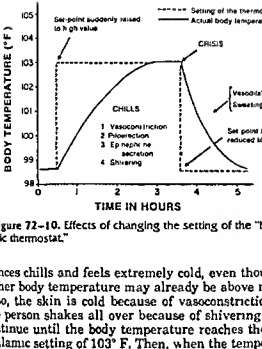


Figure 72–10. Effects of changing the setting of the "hypothalamic thermostat."

nences chills and feels extremely cold, even though his or her body temperature may already be above normal. Also, the skin is cold because of vasoconstriction, and the person shakes all over because of shivering Chills continue until the body temperature reaches the hypothalamic setting of 103° F. Then, when the temperature of the body reaches this value, the person no longer experiences chills but instead feels neither cold nor hot. As long as the factor that is causing the hypothalamic thermostat to be set at this high value continues its effect, the body temperature is regulated more or less in the normal manner but at the high temperature set-point level.

The Crisis, or "Flush." If the factor that is causing the high temperature is suddenly removed, the set-point of the hypothalamic thermostat is suddenly reduced to a lower value—perhaps even back to the normal level, as illustrated in Figure 72–10. In this instance, the body temperature is still 103° F, but the hypothalamus is attempting to regulate the temperature to 98 6° F This situation is analogous to excessive heating of the preoptic area, which causes intense sweating and sudden development of a hot skin because of vasodilatation everywhere. This sudden change of events in a febrile disease is known as the "crisis" or, more appropriately, the "flush." In olden days, before the advent of antibiotics, the crisis was always awaited, for once this occurred the doctor knew immediately that the patient's temperature would soon be falling, which is called *defervescence*.

Heat Stroke

The limits of extreme heat that one can stand depend almost entirely on whether the heat is dry or wet. If the air is completely dry and sufficient convection air currents are flowing to promote rapid evaporation from the body, a person can withstand several hours of air temperature at 150° F with no apparent ill effects On the other hand, if the air is 100 per cent humidified or if the body is in water, the body temperature begins to rise whenever the environmental temperature rises above approximately 94° F. If the person is performing very heavy work, this critical temperature level may be as low as 85° to 90° F.

Unfortunately, there is a limit to the rate at which the body can lose heat even with maximal sweating. Furthermore, when the hypothalamus becomes excessively heated, its heat-regulating ability becomes greatly depressed and sweating diminishes. As a result, a high body temperature tends to perpetuate itself unless measures are taken specifically to decrease body heat.

When the body temperature rises beyond a critical temperature, into the range of 106° to 108° F, the person is likely to develop *heat stroke*. The symptoms include dizziness, abdominal distress, sometimes delirium, and eventually loss of consciousness if the body temperature is not soon decreased. Many of these symptoms probably result from a mild degree of *circulatory shock* brought on by excessive loss of fluid and electrolytes in the sweat before the onset of symptoms. However, the hyperpyrexia itself is also exceedingly damaging to the body tissues, especially to the brain, and therefore is undoubtedly responsible for many of the effects. In fact, even a few minutes of very high body temperature can sometimes be fatal. For this reason, many authorities recommend immediate treatment of heat stroke by placing the person in an ice water bath However, because this often induces uncontrollable shivering with considerable increase in rate of heat production, others have suggested that sponge or spray cooling of the skin is likely to be more effective for rapidly decreasing the body core temperature.

Harmful Effects of the High Temperature. When the body temperature rises above 106° to 108° F, the parenchyma of many cells begins to be damaged. The pathological findings in a person who dies of hyperpyrexia are local hemorrhages and parenchymatous degeneration of cells throughout the entire body, but especially in the brain. Unfortunately, once neuronal cells are destroyed, they can never be replaced. Damage to the liver, kidneys, and other body organs can often be great enough that failure of one or more of these eventually causes death, sometimes not till several days after the heat stroke

Acclimatization to Heat. It is often extremely important to acclimatize persons to extreme heat; some examples are acclimatization of soldiers for tropical duty and acclimatization of miners for work in the 2-mile-deep gold mines of South Africa where the temperature approaches body temperature and the humidity approaches 100 per cent. Exposure of a person to heat for several hours each day while working a reasonably heavy work load will develop increased tolerance to hot and humid conditions in one to three weeks. Probably the most important physiological changes that occur during this acclimatization process are an increase in the maximum rate of sweating, increase in the plasma volume, and diminished loss of salt in the sweat and urine; the latter two of these effects result from increased secretion of aldosterone.

EXPOSURE OF THE BODY TO EXTREME COLD

Unless treated immediately, a person exposed to ice water for approximately 20 to 30 minutes ordinarily dies because of heart standstill or heart fibrillation. By that time, the internal body temperature will have fallen to about 77° F. Yet, if warmed rapidly by application of external heat, the person's life can often be saved.

Loss of Temperature Regulation at Low Temperatures. As noted in Figure 72–9, once the body temperature has fallen below 85° F the ability of the hypothalamus to regulate temperature is completely lost; it is greatly impaired even when the body temperature falls below approximately 94° F. Part of the reason

loss of temperature regulation is that the rate of heat production in each cell is depressed almost twofold for each 10° F decrease in body temperature. Also, sleepiness and even coma are likely to develop, which depress the activity of the central nervous system heat control mechanisms and prevent shivering.

Frostbite. When the body is exposed to extremely low temperatures, surface areas can actually freeze; the freezing is called *frostbite*. This occurs especially in the lobes of the ears and in the digits of the hands and feet. If the freeze has been sufficient to cause extensive formation of ice crystals in the cells, permanent damage may result, such as permanent circulatory impairment as well as local tissue damage. Often gangrene follows thawing, and the frostbitten areas are lost.

Cold-Induced Vasodilation as a Protection Against Frostbite. When the temperature of tissues falls almost to freezing, the smooth muscle in the vascular wall becomes paralyzed because of the cold itself, and sudden vasodilation occurs, often manifested by a flush of the skin. This mechanism, fortunately, helps prevent frostbite by delivering warm blood to the skin. Unfortunately, it is a far less well-developed mechanism in humans than in most lower animals that live in the cold all the time.

Artificial Hypothermia. It is very easy to decrease the temperature of a person by administering a sedative to depress the reactivity of the hypothalamic thermostat and then cooling the person with ice, cooling blankets, or otherwise until the temperature falls. The temperature can then be maintained below 90° F for several days to a week or more by continual sprinkling of cool water or alcohol on the body. Such artificial cooling is often used during heart surgery so that the heart can be stopped artificially for many minutes at a time. Cooling to this extent does not cause severe physiological results. It does slow the heart and greatly depresses body metabolism.

REFERENCES

Benzinger, T. H. Heat regulation: Homeostasis of central temperature in man. *Physiol. Rev.* *49*:671, 1969.

Bligh, J. Temperature Regulation in Mammals and Other Vertebrates. New York: American Elsevier Publishing Co., 1973.

Blix, A. S., and Steen, J. B. Temperature regulation in newborn polar homeotherms. *Physiol. Rev.* *59*:285, 1979.

Brengelmann, G. L. Circulatory adjustments to exercise and heat stress. *Annu. Rev. Physiol.* *45*:191, 1983.

Bukowiecki, L. J. Mechanisms of stimulus-calorigenesis coupling in brown adipose tissue. *Can. J. Biochem. Cell. Biol.*, *62*:623, 1984.

Cabanac, M. Thermoregulatory behavior. *In* MTP International Review of Science: Physiology, Vol. 7. Baltimore, University Park Press, 1974, p. 231.

Cabanac, M. Temperature regulation. *Annu. Rev. Physiol.*, *37*:415, 1975.

Calder, W. A., III. Scaling of physiological processes in homeothermic animals. *Annu. Rev. Physiol.*, *43*:301, 1981.

Cena, K., and Clark, J. A. Transfer of heat through animal coats and clothing. *In* Robertshaw, D. (ed.), International Review of Physiology: Environmental Physiology III, Vol. 20. Baltimore, University Park Press, 1979, p. 1.

Crawshaw, L. I. Temperature regulation in vertebrates. *Annu. Rev. Physiol.*, *42*:473, 1980.

Elizondo, R. Temperature regulation in primates. *Int. Rev. Physiol.*, *15*:71, 1977.

Galanter, E. Detection and discrimination of environmental change. *In* Darian-Smith, I. (ed.), Handbook of Physiology, Sec. 1, Vol. III. Bethesda: American Physiological Society, 1984, p. 103.

Hales, J. E. (ed.). Thermal Physiology. New York, Raven Press, 1984.

Hales, J. R. S. Physiological responses to heat. *In* MTP International Review of Science: Physiology, Vol. 7. Baltimore, University Park Press, 1974, p. 107.

Hardy, J. D. Physiology of temperature regulation. *Physiol. Rev.* *41*:521, 1961.

Hardy, R. N. Temperature and Animal Life. Baltimore, University Park Press, 1979.

Harrison, M. H. Effects of thermal stress and exercise on blood volume in humans. *Physiol. Rev.*, *65*:149, 1985.

Heller, H. C., *et al.* The thermostat of vertebrate animals. *Sci. Am.* *239*:(2):102, 1978.

Hellon, R. Thermoreceptors. *In* Shepherd, J. T., and Abboud, F. M. (eds.), Handbook of Physiology, Sec. 1, Vol. III. Bethesda, American Physiological Society, 1983, p. 659.

Hensel, H. Neural processes in thermoregulation. *Physiol. Rev.*, *53*:94, 1973.

Hensel, H. Thermoreceptors. *Annu. Rev. Physiol.*, *36*:233, 1974.

Himms-Hagen, J. Cellular thermogenesis. *Annu. Rev. Physiol.*, *39*:31, 1976.

Jansky, L. Non-shivering thermogenesis and its thermoregulatory significance. *Biol. Rev.* *48*:85, 1973.

Kluger, M. J. Temperature regulation, fever, and disease. *In* Robertshaw, D. (ed.), International Review of Physiology: Environmental Physiology III, Vol. 20. Baltimore, University Park Press, 1979, 209.

Maclean, D., and Emslie-Smith, D. Accidental Hypothermia. Oxford: Blackwell Scientific Publications, 1977.

Mitchell, D. Physical basis of thermoregulation. *Int. Rev. Physiol.*, *15*, 1977.

Nicholls, D. G., and Locke, R. M. Thermogenic mechanisms in brown fat. *Physiol. Rev.* *64*:1, 1984.

Precht, H., *et al.* Temperature and Life. New York, Springer-Verlag, 1973.

Prosser, C. L., and Nelson, D. O. The role of nervous systems in temperature adaptation of poikilotherms. *Annu. Rev. Physiol.*, *43*:281, 1981.

Quinton, P. M. Sweating and its disorders. *Annu. Rev. Med.*, *34*:45, 1983.

Reeves, R. R. The interaction of body temperature and acid-base balance in ectothermic vertebrates. *Annu. Rev. Physiol.*, *39*:559, 1977.

Robertshaw, D. Role of the adrenal medulla in thermoregulation. *Int. Rev. Physiol.* *15*:189, 1977.

Rowell, L. B. Cardiovascular adjustments to thermal stress. *In* Shepherd, J. T., and Abboud, F. M. (eds.), Handbook of Physiology, Sec. 1, Vol. III. Bethesda, American Physiological Society, 1983, p. 967.

Satinoff, E. (ed.). Thermoregulation. Stroudsburg, Pa., Dowden, Hutchinson & Ross, New York, Academic Press, 1979.

Schmidt-Nielsen, K. Animal Physiology: Adaptation and Environment. London, Cambridge University Press, 1975.

Sloan, A. W. Man in Extreme Environments. Springfield, Ill., Charles C Thomas, 1979.

Svanes, K. Effects of temperature on blood flow. *In* Kaley, G., and Altura, B. M. (eds.), Microcirculation, Vol. III. Baltimore, University Park Press, 1977.

Thompson, G. E. Physiological effects of cold exposure. *Int. Rev. Physiol.*, *15*:29, 1977.

Underwood, L. S., and Tiezzen, L. L. (eds.), Comparative Mechanisms of Cold Adaptation. New York, Academic Press, 1979.

Wang, L. C. H., and Hudson, J. W. (eds.), Strategies in Cold: Natural Torpidity and Thermogenesis. New York, Academic Press, 1978.

Webster, A. J. F. Physiological effects of cold exposure. *In* MTP International Review of Science: Physiology, Vol. 7. Baltimore, University Park Press, 1974, p. 33.

Wyndham, C. H. The physiology of exercise under heat stress. *Annu. Rev. Physiol.*, *35*:193, 1973.

73

Dietary Balances; Regulation of Feeding; Obesity and Starvation; Vitamins and Minerals

The intake of food must always be sufficient to supply the metabolic needs of the body and yet not so much as to cause obesity. Also, since different foods contain different proportions of proteins, carbohydrates, and fats, appropriate balance must be maintained among these different types of food so that all segments of the body's metabolic systems can be supplied with the requisite materials. This chapter, therefore, discusses the problems of balance among the three major types of food and also the mechanisms by which the intake of food is regulated in accordance with the metabolic needs of the body.

DIETARY BALANCES

ENERGY AVAILABLE IN FOODS

The energy liberated from each gram of carbohydrate as it is oxidized to carbon dioxide and water is 4.1 Calories, and that liberated from fat is 9.3 Calories. The energy liberated from metabolism of the average protein of the diet as each gram is oxidized to carbon dioxide, water, and urea is 4.35 Calories. Also, these different substances vary in the average percentages that are absorbed from the gastrointestinal tract: approximately 98 per cent of the carbohydrate, 95 per cent of the fat, and 92 per cent of the protein. Therefore, in round figures the average *physiologically available energy* in each gram of the three different foodstuffs in the diet is.

	Calories
Carbohydrates	4.0
Fat	9.0
Protein	4.0

Average Americans receive approximately 15 per cent of their energy from protein, about 40 per cent from fat, and 45 per cent from carbohydrates. In most other parts of the world the quantity of energy derived from carbohydrates far exceeds that derived from both proteins and fats. Indeed, in Mongolia the energy received from fats and proteins combined is said to be no greater than 15 to 20 per cent.

Table 73–1 gives the compositions of selected foods,

Table 73–1 PROTEIN, FAT, AND CARBOHYDRATE CONTENT OF DIFFERENT FOODS

Food	Protein %	Fat %	Carbohydrate %	Fuel Value per 100 Grams, Calories
Apples	0.3	0.4	14.9	64
Asparagus	2.2	0.2	3.9	26
Bacon, fat broiled	6.2	76.0	0.7	712
	25.0	55.0	1.0	599
Beef, medium	17.5	22.0	1.0	268
Beets, fresh	1.6	0.1	9.6	46
Bread, white	9.0	3.6	49.8	268
Butter	0.6	81.0	0.4	733
Cabbage	1.4	0.2	5.3	29
Carrots	1.2	0.3	9.3	45
Cashew nuts	19.6	47.2	26.4	609
Cheese, Cheddar, American	23.9	32.3	1.7	393
Chicken, total edible	21.6	2.7	1.0	111
Chocolate	(5.5)	52.9	(18.0)	570
Corn (maize), entire	10.0	4.3	73.4	372
Haddock	17.2	0.3	0.5	72
Lamb, leg, intermediate	18.0	17.5	1.0	230
Milk, fresh whole	3.5	3.9	4.9	69
Molasses, medium	0.0	0.0	(60.0)	240
Oatmeal, dry, uncooked	14.2	7.4	68.2	396
Oranges	0.9	0.2	11.2	50
Peanuts	26.9	44.2	23.6	600
Peas, fresh	6.7	0.4	17.7	101
Pork, ham, medium	15.2	31.0	1.0	340
Potatoes	2.0	0.1	19.1	85
Spinach	2.3	0.3	3.2	25
Strawberries	0.8	0.6	8.1	41
Tomatoes	1.0	0.3	4.0	23
Tuna, canned	24.2	10.8	0.5	164
Walnuts, English	15.0	64.4	15.6	702

illustrating especially the high proportions of fats and proteins in meat products and the high proportions of carbohydrates in most vegetable and grain products. Fat is deceptive in the diet, for it usually exists as 100 per cent fat, whereas both proteins and carbohydrates are mixed in watery media so that each of these normally represents less than 25 per cent of the weight of the food. Therefore, the fat of one pat of butter mixed with an entire helping of potato often contains as much energy as all the potato itself.

Daily Requirement for Protein. Twenty to 30 grams of the body proteins are degraded and used for producing other body chemicals daily. Therefore, all cells must continue to form new proteins to take the place of those that are being destroyed, and a supply of protein is needed in the diet for this purpose. An average person can maintain normal stores of protein provided the *daily intake is above 30 to 55 grams*.

However, as was discussed in Chapter 69, some proteins have inadequate quantities of certain essential amino acids and therefore cannot be used to form the body proteins. Such proteins are called *partial proteins*, and when they are present in large quantities in the diet, the daily requirement of protein is much greater than normally. In general, proteins derived from animal foodstuffs are more nearly complete than are proteins derived from vegetable and grain sources. A special example of dietary deficiency resulting from partial proteins occurs in the diet of many African natives who subsist primarily on corn meal. The protein of corn has almost no tryptophan, therefore, for practical purposes, the entire diet of these natives is almost completely protein deficient. As a result, these natives, especially the children, develop the protein deficiency syndrome called *kwashiorkor*, which consists of failure to grow, lethargy, depressed mentality, and hypoprotein edema.

Carbohydrates and Fats as "Protein Sparers." It was also noted in Chapters 67 and 68 that when the diet contains an abundance of carbohydrates and fats, almost all the body's energy is derived from these two substances and very little is derived from proteins. Therefore, both carbohydrates and fats are said to be *protein sparers*. On the other hand, in starvation, after the carbohydrates and fats have been depleted the body's protein stores are then consumed rapidly for energy, sometimes at rates approaching several hundred grams per day rather than at the normal daily rate of 30 to 55 grams.

Clinical and Experimental Methods for Determining Metabolic Utilization of Proteins, Carbohydrates, and Fats

In animal experiments, and also in human beings who have metabolic problems, it is often important to determine the relative rates of utilization of proteins, carbohydrates, and fats. Two principal methods are used for this purpose.

Determination of the Rate of Protein Metabolism in the Body. The average protein of the diet contains approximately 16 per cent nitrogen, and the remaining 84 per cent is composed almost entirely of carbon, hydrogen, and oxygen. During metabolism of the protein about 90 per cent of the protein nitrogen is excreted in the urine in the form of urea, uric acid, creatinine, and other less important nitrogen products. The remaining 10 per cent is excreted in the feces. Therefore, it is a simple matter to estimate quite accurately the rate of protein breakdown in the body by simply measuring the amount of nitrogen in the urine, then adding another 11 per cent for the nitrogen excreted in the feces, and finally multiplying by 6.25 to determine the total amount of protein metabolism in grams per day. The excretion of 8 grams of nitrogen in the urine each day calculates to be about 55 grams of protein breakdown.

If one also knows the daily intake of protein, the so called *nitrogen balance* can then be calculated. If the daily intake of protein is less than the daily breakdown of protein, the person is said to have a *negative* nitrogen balance, which means that his or her body stores of protein are decreasing daily.

Relative Utilization of Fat and Carbohydrates—The "Respiratory Quotient." When carbohydrates are metabolized with oxygen, for each molecule of oxygen consumed exactly 1 carbon dioxide molecule is formed. This ratio of carbon dioxide output to oxygen usage is called the *respiratory quotient*, and therefore the respiratory quotient for carbohydrates is 1.0.

On the other hand, when fat is oxidized in the body cells, an average of only 70 carbon dioxide molecules are formed for each 100 molecules of oxygen consumed. Therefore, the respiratory quotient for the metabolism of fat averages 0.70. And, in a similar manner, when proteins are oxidized by the cells, the average respiratory quotient is 0.80. The reason for these lower respiratory quotients for fat and protein than for carbohydrates is that a large share of the oxygen metabolized with these foods is required to combine with the excess hydrogen ions present in them instead of forming carbon dioxide.

Now let us see how one can utilize this respiratory quotient to determine the relative utilization of different foods by the body. First, it will be recalled from Chapter 40 that the output of carbon dioxide by the lungs divided by the uptake of oxygen during the same period of time is called the *respiratory exchange ratio*, and this can be determined by measuring the pulmonary exchanges of carbon dioxide and oxygen. Over a period of an hour or more, the respiratory exchange ratio exactly equals the average respiratory quotient of the metabolic reactions throughout the body. Thus, when the respiratory exchange ratio measured over such a period of time is 0.95, one can also assume that the respiratory quotient is almost exactly 0.95 as well. And it is obvious that if a person has a respiratory quotient of 1.0, he can only be metabolizing carbohydrates, because the respiratory quotients for both fat and protein metabolism are considerably less than 1.0. Likewise, when the respiratory quotient is approximately 0.70, the body is metabolizing fat almost entirely, to the exclusion of carbohydrates and proteins. And, finally, if we ignore the normally small amount of protein metabolism, respiratory quotients between 0.70 and 1.0 describe the approximate ratios of carbohydrate to fat metabolism. To be more exact, though, one can first determine the protein utilization by measuring nitrogen excretion, and then, using appropriate mathematical formulation, one can calculate almost exactly the utilization of the three different foodstuffs.

Some of the important findings from studies of respiratory quotient are the following:

(1) Immediately after a meal almost all the food that is metabolized is carbohydrates, so that the respiratory quotient at that time approaches 1.0.

(2) Approximately 8 to 10 hours following a meal, the body has already used up most of its stores of carbohydrate, and the respiratory quotient approaches that for fat metabolism, approximately 0.70.

(3) In diabetes mellitus, very little carbohydrate can be utilized by the body's cells under any conditions because insulin is required for this. Therefore, when diabetes is very severe, the respiratory quotient remains most of the time very near to that for fat metabolism, 0.70.

REGULATION OF FOOD INTAKE

Hunger. The term hunger means a craving for food, and it is associated with a number of objective sensations. For instance, in a person who has not had food for many hours, the stomach undergoes intense rhythmic contractions called *hunger contractions*. These cause a tight or gnawing feeling in the pit of the stomach and sometimes actually cause pain called *hunger pangs*. In addition to the hunger pangs, the hungry person also becomes more tense and restless than usual.

Some physiologists actually define hunger as the tonic contractions of the stomach. However, even after the stomach is completely removed, the psychic sensations of hunger still occur, and craving for food still makes the person search for an adequste food supply.

Appetite. The term appetite is often used in the same sense as hunger except that it usually implies desire for specific types of food instead of food in general. Therefore, appetite helps a person choose the quality of food to eat.

Satiety. Satiety is the opposite of hunger. It means a feeling of fulfillment in the quest for food. Satiety usually results from a filling meal, particularly when the person's nutritional storage depots, the adipose tissue and the glycogen stores, are already filled.

NEURAL CENTERS FOR REGULATION OF FOOD INTAKE

Hunger and Satiety Centers. Stimulation of the *lateral hypothalamus* causes an animal to eat voraciously, which is called *hyperphagia*. On the other hand, stimulation of the *ventromedial nuclei of the hypothalamus* causes complete satiety, and, even in the presence of highly appetizing food, the animal will still refuse to eat, which is *aphagia*. Conversely, destructive lesions of the two respective areas cause results exactly opposite to those caused by stimulation. That is, ventromedial lesions cause voracious and continued eating until the animal becomes extremely obese, sometimes as large as four times normal size. And lesions of the lateral hypothalamic nuclei cause complete lack of desire for food and progressive inanition of the animal. Therefore, we can label the lateral nuclei of the hypothalamus as a *hunger center* or *feeding center*, and we can label the ventromedial nuclei of the hypothalamus as a *satiety center*.

The feeding center operates by directly exciting the emotional drive to search for food (while also stimulating other emotional drives as well—see Chapters 56 and 57). On the other hand, it is believed that the satiety center operates primarily by inhibiting the feeding center.

Other Neural Centers That Enter into Feeding. If the brain is sectioned below the hypothalamus but above the mesencephalon, the animal can still perform the basic mechanical features of the feeding process. It can salivate, lick its lips, chew food, and swallow. Therefore, *the actual mechanics of feeding are controlled by centers in the* brain stem. The function of the hypothalamus in feeding, then, is to control the quantity of food intake and to excite the lower centers to activity.

Higher centers than the hypothalamus also play important roles in the control of feeding, particularly in the control of appetite. These centers include especially the *amygdala* and the *cortical areas of the limbic system*, all of which are closely coupled with the hypothalamus. It will be recalled from the discussion of the sense of smell that portions of the amygdala are a major part of the olfactory nervous system. Destructive lesions in the amygdala have demonstrated that some of its areas greatly increase feeding, while others inhibit feeding. In addition, stimulation of some areas of the amygdala elicits the mechanical act of feeding. However, the most important effect of destruction of the amygdala on both sides of the brain is a "psychic blindness" in the choice of foods. In other words, the animal (and presumably the human being as well) loses or at least partially loses the mechanism of appetite control of type and quality of food that it eats.

The cortical regions of the limbic system, including the infraorbital regions, the hippocampal gyrus, and the cingulate gyrus, all have areas that when stimulated can either increase or decrease feeding activities. These areas seem especially to play a role in the animal's drive to search for food when it is hungry. It is presumed that these centers are also responsible, probably operating in association with the amygdala and hypothalamus, for determining the quality of food that is eaten. For instance, a previous unpleasant experience with almost any type of food often destroys a person's appetite for that food thenceforth.

FACTORS THAT REGULATE FOOD INTAKE

We can divide the regulation of food into (1) *nutritional regulation*, which is concerned primarily with maintenance of normal quantities of nutrient stores in the body, and (2) *alimentary regulation*, which is concerned primarily with the immediate effects of feeding on the alimentary

tract and is sometimes called *peripheral regulation* or *short-term regulation*.

Nutritional Regulation. An animal that has been starved for a long time and is then presented with unlimited food eats a far greater quantity than does an animal that has been on a regular diet. Conversely, an animal that has been force-fed for several weeks eats little when allowed to eat according to its own desires. Thus, the feeding center in the hypothalamus is geared to the nutritional status of the body. Some of the nutritional factors that control the degree of activity of the feeding center are the following.

Availability of Glucose to the Body Cells—The Glucostatic Theory of Hunger and of Feeding Regulation. It has long been known that a decrease in blood glucose concentration is associated with development of hunger, which has led to the so-called *glucostatic theory of hunger and of feeding regulation*. When the blood glucose level falls too low, this automatically causes the animal to increase its feeding, which eventually returns the glucose concentration back toward normal. Two other observations also support the glucostatic theory: (1) An increase in blood glucose level increasea tha measured electrical activity in the satiety center in the ventromedial nuclei of the hypothalamus and simultaneously decreases the alectrical activity in the feeding center of the lateral nuclei (2) Chemical studies show that the vantromedial nuclei (the satiety center) concentrate glucose while other areas of the hypothalamus fail to concentrate glucose; therefore, it is assumed that glucose acts by increasing the degree of satiety.

Effect of Blood Amino Acid Concentration on Feeding. An increase in amino acid concentration in the blood also reduces feeding, and a decrease enhances feeding. In general, though, this effect is not as powerful as the glucostatic mechanism.

Effect of Fat Metabolites on Feeding—Long-Term Regulation. The overall degree of feeding varies almost inversely with the amount of adipose tissue in the body. That is, as the quantity of adipose tissue increases, the rate of feeding decreases. Therefore, many physiologists believe that *long-term regulation* of feeding is controlled mainly by fat metabolites of undiscovered nature. This is called the "lipostatic" theory of feeding regulation. In support of this is the fact that the long-term average concentration of free fatty acids in the blood is directly proportional to the quantity of adipose tissue in the body. Therefore, it is likely that the free fatty acids or some other similar fat metabolites act in the same manner as glucose and amino acids to cause a negative feedback regulatory effect on feeding. It is also possible, if not probable, that this is the most important long-term regulator of feeding.

Interrelationship Between Body Tempera- and Food Intake. When an animal is exposed to cold, it tends to overeat; when exposed to heat, it tends to undereat. This is caused by interaction within the hypothalamus between the temperature-regulating system (see Chapter 72) and the food intake-regulating system. It is important because increased food intake in the cold animal (a) increases its metabolic rate and (b) provides increased fat for insulation, both of which tend to correct the cold state

Summary of Long-Term Regulation. Even though our information on the different feedback factors in long-term feeding regulation is imprecise, we can make the following general statement: When the nutrient stores of the body fall below normal, the feeding center of the hypothalamus becomes highly octive, and the person exhibits increased hunger; on the other hand, when the nutrient stores are abundant, the person loses the hunger and develops a state of satiety.

Alimentary Regulation (Short-Term, Nonmetabolic Regulation). The degree of hunger or satiety at different times of the day depends to a great extent on habit For instance, normally, people have the habit of eating three meals a day, and, if they miss one, they are likely to develop a state of hunger at mealtime despite completely odequate nutritional stores in their tissues. But, in addition to habit, several other short-term physiological stimuli—mainly related to the alimentary tract—can alter one's desire for food for several hours at a time, as follows:

Gastrointestinal Filling. When the gastrointestinal tract becomes distended, especially the stomach or the duodenum, inhibitory signals temporarily suppress the feeding center, thereby reducing the desire for food. This effect probably depends mainly on sensory signals transmitted through the vagi, but part of the effect still persists after the vagi and the sympathetic nerves from the upper gastrointestinal tract have been severed. Therefore, somatic sensory signals from the stretched abdomen might also play a role. And, recently it has been found that hormonal feedback also suppresses feeding, for *cholecystokinin* that is released in response mainly to fat entering the duodenum has o strong effect on inhibition of further eating

Obviously, these mechanisms are of particular importance in bringing one's feeding to a halt during a heavy meal.

Metering of Food by Head Receptors. When a person with on esophageal fistula is fed large quantities of food, even though this food is immediately lost again to the exterior the degree of hunger is decreased after a reasonable quantity of food has passed through the mouth. This effect occurs despite the fact that the gastrointestinal tract does not become the least bit filled. Therefore, it is postulated that various "head factors" relating to feeding, such as chewing, salivation, swallowing, and tasting, "meter" the food as it passes through the mouth, and after a certain amount

has passed through, the hypothalamic feeding center becomes inhibited. However, the inhibition caused by this metering mechanism is considerably less intense and less lasting, usually lasting only 20 to 40 minutes, than is the inhibition caused by gastrointestinal filling.

Importance of Having Both Long- and Short-term Regulatory Systems for Feeding. The long-term regulatory system, especially the lipostatic feedback mechanism, obviously helps an animal maintain constant stores of nutrients in its tissues, preventing these from becoming too low or too high. On the other hand, the short-term regulatory stimuli make the animal feed only when the gastrointestinal tract is receptive to food. Thus, food passes through its gastrointestinal tract fairly continuously so that its digestive, absorptive, and storage mechanisms can all work at a steady pace rather than only when the animal needs food for energy.

OBESITY

Energy Input Versus Energy Output. When greater quantities of energy (in the form of food) enter the body than are expended, the body weight increases. Therefore, obesity is obviously caused by excess energy input over energy output. For each 9.3 Calories of excess energy entering the body, 1 gram of fat is stored.

Excess energy input occurs *only during the developing phase of obesity*, and once a person has become obese, all that is required to remain obese is that the energy input equal the energy output. For the person to reduce in weight, the input must be *less* than the output. Indeed, studies of obese persons have shown that the intake of food of most of them in the static stage of obesity (after the obesity has already been attained) is approximately the same as that for normal persons.

Effect of Muscular Activity on Energy Output. About one third of the energy used each day by the normal person goes into muscular activity, and in the laborer as much as two thirds or occasionally three fourths is used in this way. Since muscular activity is by far the most important means by which energy is expended in the body, it is frequently said that obesity results from *too high a ratio of food intake to daily exercise.*

Abnormal Feeding Regulation as a Pathological Cause of Obesity

We have already emphasized that the rate of feeding is normally regulated in proportion to the nutrient stores in the body. When these stores begin to approach an optimal level in a normal person, feeding is automatically reduced to prevent overstorage. However, in many obese persons this is not true, for feeding does not slacken until body weight is far above normal. Therefore, in effect, obesity is often caused by an abnormality of the feeding regulatory mechanism. This can result from either psychogenic factors that affect the regulation or actual abnormalities of the hypothalamus itself.

Psychogenic Obesity. Studies of obese patients show that a large proportion of obesity results *from psychogenic factors.* Perhaps the most common psychogenic factor contributing to obesity is the prevalent idea that healthy eating habits require three meals a day and that each meal must be filling. Many children are forced into this habit by overly solicitous parents, and the children continue to practice it throughout life. In addition, persons are known often to gain large amounts of weight during or following stressful situations, such as the death of a parent, a severe illness, or even mental depression. It seems that eating is often a means of release from tension.

Hypothalamic Abnormalities as a Cause of Obesity. In the preceding discussion of feeding regulation, it was pointed out that lesions in the ventromedial nuclei of the hypothalamus cause an animal to eat excessively and become obese. It has also been discovered that such lesions are associated with excess insulin production, which in turn increases fat deposition. Also, many persons with hypophysial tumors that encroach on the hypothalamus develop progressive obesity, illustrating that obesity in the human being, too, can definitely result from damage to the hypothalamus.

Yet, in the normal obese person hypothalamic damage is almost never found. Nevertheless, it is possible that the functional organization of the feeding center is different in the obese person from that of the nonobese person. For instance, a normally obese person who has reduced to normal weight by strict dietary measures usually develops hunger that is demonstrably far greater than that of the normal person. This indicates that the "set-point" of the obese person's feeding center is at a much higher level of nutrient storage than that of the normal person.

Genetic Factors in Obesity. Obesity definitely runs in families. Furthermore, identical twins usually maintain weight levels within 2 pounds of each other throughout life if they live under similar conditions, or within 5 pounds of each other if their conditions of life differ markedly. This might result partly from eating habits engendered during childhood, but it is generally believed that this close similarity between twins is genetically controlled.

The genes can direct the degree of feeding in several different ways, including (1) a genetic abnormality of the feeding center to set the level of nutrient storage high or low and (2) abnormal hereditary psychic factors that either whet the appetite or cause the person to eat as a "release" mechanism.

Genetic abnormalities in the *chemistry of fat storage* are also known to cause obesity in certain strains of rats and mice. In one strain of rats, fat is easily stored in the adipose tissue, but the quantity of hormone-sensitive lipase in the adipose tissue is greatly reduced, so that little of the fat can be removed. In addition, these rats develop hyperinsulinism which also promotes fat storage. This combination obviously results in a one-way path, the fat continually being deposited but never released. In a strain of obese mice that has been studied, there is excess of fatty acid synthetase, which causes excess synthesis of fatty acids. Thus, similar genetic mechanisms are other possible causes of obesity in human beings.

Childhood Overnutrition as a Possible Cause of Obesity. The rate of formation of new fat cells is especially rapid in the first few years of life, and the greater the rate of fat storage the greater also becomes the number of fat cells. In obese children the number of fat cells is often as much as three times that in normal

children. However, after adolescence, the number of fat cells remains almost identically the same throughout the remainder of life. Therefore, it has been suggested that overfeeding children, especially in infancy and to a lesser extent during the older years of childhood, can lead to a lifetime of obesity. The person who has excess fat cells is thought to have a higher setting of fat storage by the hypothalamic feedback autoregulatory mechanism for control of adipose tissues

In less obese persons, especially those who become obese in middle or old age, most of the obesity results from hypertrophy of already existing fat cells This type of obesity is far more susceptible to treatment than is the lifelong type An interesting peculiarity of this type of obesity is that these persons have excess secretion of insulin, similar to the excess insulin secretion that occurs in animals with lesions in the ventromedial nuclei of the hypothalamus

TREATMENT OF OBESITY

Treatment of obesity depends simply on decreasing energy input below energy expenditure In other words, this means partial starvation For this purpose, most diets are designed to contain large quantities of "bulk" which, in general, are made up of non-nutritive cellulose substances This bulk distends the stomach and thereby partially appeases the hunger In most lower animals such a procedure simply makes the animal increase its food intake still further, but human beings can often fool themselves because their food intake is sometimes controlled as much by habit as by hunger As pointed out later in connection with starvation, it is important to prevent vitamin deficiencies during the dieting period

Various *drugs for decreasing the degree of hunger* have been used in the treatment of obesity. The most important of these is *amphetamine* (or amphetamine derivatives), which directly inhibits the feeding center in the lateral nuclei of the hypothalamus However, there is danger in using this drug because it simultaneously overexcites the central nervous system, making the person nervous and elevating the blood pressure Also, a person soon adapts to the drug so that weight reduction is usually no greater than 10 per cent

Finally, the more exercise one takes, the greater is the daily energy expenditure and the more rapidly the obesity disappears Therefore, forced exercise is often an essential part of the treatment for obesity

INANITION

Inanition is the exact opposite of obesity. In addition to inanition caused by inadequate availability of food, both psychogenic and hypothalamic abnormalities can on occasion cause greatly decreased feeding. One such condition, *anorexia nervosa,* is an abnormal psychic state in which a person loses all desire for food and even becomes nauseated by food; as a result, severe inanition occurs Also, destructive lesions of the hypothalamus, often caused by vascular thrombosis, frequently cause a condition called *cachexia;* the term simply means severe inanition

STARVATION

Depletion of Food Stores in the Body Tissues Starvation. Even though the tissues preferentially use carbohydrate for energy over both fat and protein, the quantity of carbohydrate stores of the body is only a few hundred grams (mainly glycogen in the liver and muscles), and it can supply the energy required for body function for perhaps half a day. Therefore, except for the first few hours of starvation, the major effects are progressive depletion of tissue fat and protein Since fat is the prime source of energy, its rate of depletion continues unabated, as illustrated in Figure 73–1, until most of the fat stores in the body are gone

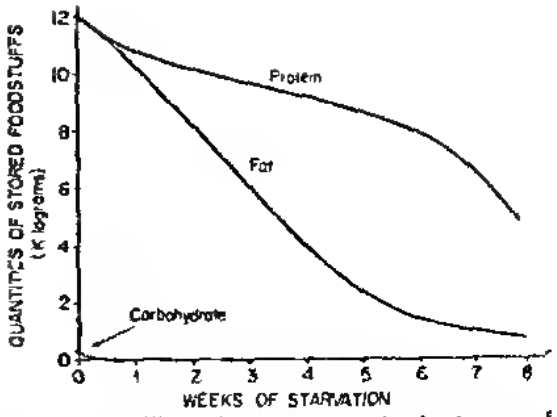


Figure 73–1. Effect of starvation on the food stores of the body

Protein undergoes three different phases of depletion rapid depletion at first, then greatly slowed depletion and finally, rapid depletion again shortly before death The initial rapid depletion is caused by conversion of protein to glucose in the liver by the process of gluconeogenesis About two thirds of the glucose thus formed is used mainly to supply energy to the brain which under normal circumstances, utilizes almost no other metabolic substrate for energy besides glucose However after the readily mobilizable protein stores have been depleted during the early phase of starvation, the remaining protein is not so easily removed from the tissues At this time, the rate of gluconeogenesis decreases to one third to one fifth its previous rate, and the rate of depletion of protein becomes greatly decreased The lessened availability of glucose then initiates a series of events that often lead to *ketosis,* which was discussed in Chapter 68 Fortunately, the ketone bodies, like glucose, can cross the blood-brain barrier and can be utilized by the brain cells for energy. Therefore, approximately two thirds of the brain's energy now is derived from these ketone bodies, principally from beta-hydroxybutyrate This sequence of events thus leads to at least partial preservation of the protein stores of the body

However, there finally comes a time when the fat stores also are almost totally depleted, and the only remaining source of energy is proteins At that time the protein stores once again enter a stage of rapid depletion Since the proteins are essential for maintenance of cellular function, death ordinarily ensues when the proteins of the body have been depleted to approximately one half their normal level.

Vitamin Deficiencies in Starvation. The stores of some of the vitamins, especially the water-soluble vitamins—the vitamin B group and vitamin C—do not last long during starvation. Consequently, after a week or more of starvation mild vitamin deficiencies usually begin to appear, and after several weeks severe vitamin deficiencies can occur. Obviously, these can add to the debility that leads to death.

Table 73–2 REQUIRED DAILY AMOUNTS OF THE VITAMINS

A	5000 IU
Thiamine	1.5 mg
Riboflavin	1.8 mg
Niacin	20 mg
Ascorbic acid	45 mg
D	400 IU
E	15 IU
K	none
Folic acid	0.4 mg
B_{12}	3 μg
Pyridoxine	2 mg
Pantothenic acid	unknown

IU, International units.

ITAMINS

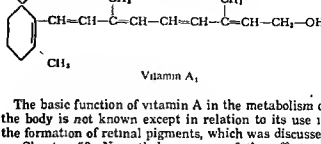
Daily Requirements of Vitamins. A vitamin is an rganic compound needed in small quantities for oper- ion of normal bodily metabolism and that cannot be anufactured in the cells of the body When lacking in e diet, vitamins can cause specific metabolic deficits Table 73–2 lists the amounts of important vitamins quired daily by the average adult man These require- ents vary considerably, depending on such factors as) body size, (b) rate of growth, (c) amount of exercise,) disease and fever, and (e) special need for vitamin D pregnant or lactating women Also, a number of etabolic deficits occur pathologically in which the tamins themselves cannot be utilized properly in the ody; in such conditions the requirement for one or more ecific vitamins may be extreme

Storage of Vitamins in the Body. Vitamins are ored to a slight extent in all the cells However, some itamins are stored to a major extent in the liver For istance, the quantity of vitamin A stored in the liver ay be sufficient to maintain a person without any itake of vitamin A for up to ten months, and ordinarily e quantity of vitamin D stored in the liver is sufficient maintain a person for two to four months without ny additional intake of vitamin D

The storage of most water-soluble vitamins is rela- ively slight; this applies especially to the vitamin B ompounds, for when a person's diet is deficient in itamin B compounds, clinical symptoms of the defi- iency can sometimes be recognized within a few days xcept for vitamin B_{12} which can last in the liver for a ear or longer). Absence of vitamin C, another water- oluble vitamin, can cause symptoms within a few weeks nd can cause death from scurvy in 20 to 30 weeks

ITAMIN A

Vitamin A occurs in animal tissues as *retinol*, the ormula of which is illustrated. This vitamin does not ccur in foods of vegetable origin, but *provitamins* for he formation of vitamin A do occur in abundance in any vegetable foods. These are the yellow and red *arotenoid pigments* which, since they have chemical tructures similar to that of vitamin A, can be changed nto vitamin A in the liver.



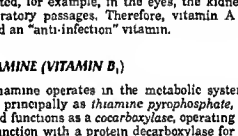
Vitamin A_1

The basic function of vitamin A in the metabolism of the body is not known except in relation to its use in the formation of retinal pigments, which was discussed in Chapter 59. Nevertheless, some of the effects of vitamin A lack have been well documented. In addition to the need for vitamin A to form the visual pigments and therefore to prevent night blindness, it is also necessary for normal growth of most cells of the body and especially for normal growth and proliferation of the different types of epithelial cells When vitamin A is lacking, the epithelial structures of the body tend to become stratified and keratinized. Vitamin A deficiency manifests itself by (1) scaliness of the skin and sometimes acne, (2) failure of growth of young animals, including cessation of skeletal growth, (3) failure of reproduction, associated especially with atrophy of the germinal epithelium of the testes and sometimes with interruption of the female sexual cycle, and (4) keratinization of the cornea with resultant corneal opacity and blindness

Also, the damaged epithelial structures often become infected, for example, in the eyes, the kidneys, or the respiratory passages. Therefore, vitamin A has been called an "anti-infection" vitamin.

THIAMINE (VITAMIN B_1)

Thiamine operates in the metabolic systems of the body principally as *thiamine pyrophosphate*, this compound functions as a *cocarboxylase*, operating mainly in conjunction with a protein decarboxylase for decarboxylation of pyruvic acid and other α-keto acids, as discussed in Chapter 67.



Thiamine chloride

Thiamine deficiency causes decreased utilization of pyruvic acid and some amino acids by the tissues but increased utilization of fats Thus, thiamine is specifically needed for final metabolism of carbohydrates and many amino acids. Probably the decreased utilization of these nutrients is the responsible factor for many debilities associated with thiamine deficiency.

Thiamine Deficiency and the Nervous System. The central nervous system depends almost entirely on the metabolism of carbohydrates for its energy. In thiamine deficiency the utilization of glucose by nervous tissue may be decreased as much as 50 to 60 per cent.

The neuronal cells of the central nervous system frequently show chromatolysis and swelling during thiamine deficiency, changes that are characteristic of neuronal cells with poor nutrition. Obviously, such changes as these can disrupt communication in many different portions of the central nervous system.

Also, thiamine deficiency can cause *degeneration of myelin sheaths* of nerve fibers both in the peripheral nerves and in the central nervous system. The lesions in the peripheral nerves frequently cause these nerves to become extremely irritable, resulting in "polyneuritis" characterized by pain radiating along the course of one or more peripheral nerve fibers, and fiber tracts in the cord can degenerate to such an extent that *paralysis* occasionally results, and, even in the absence of paralysis, the muscles atrophy, with resultant severe weakness.

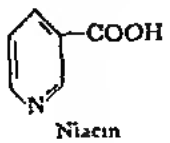
Thiamine Deficiency and the Cardiovascular System. Thiamine deficiency also weakens the heart muscle, so that a person with severe thiamine deficiency sometimes develops *cardiac failure*. Furthermore, the return of blood to the heart may be increased to as much as two times normal. This occurs because thiamine deficiency causes *peripheral vasodilation* throughout the circulatory system, possibly as a result of metabolic deficiency in the smooth muscle of the vascular system itself. Therefore, the cardiac effects of thiamine deficiency are due partly to excessive return of blood to the heart and partly to primary weakness of the cardiac muscle. *Peripheral edema* and *ascites* also occur to a major extent in some persons with thiamine deficiency because of the cardiac failure.

Thiamine Deficiency and the Gastrointestinal Tract. Among the gastrointestinal symptoms of thiamine deficiency are indigestion, severe constipation, anorexia, gastric atony, and hypochlorhydria. All these effects possibly result from failure of the smooth muscle and glands of the gastrointestinal tract to derive sufficient energy from carbohydrate metabolism.

The overall picture of thiamine deficiency, including polyneuritis, cardiovascular symptoms, and gastrointestinal disorders, is frequently referred to as "beriberi"—especially when the cardiovascular symptoms predominate.

NIACIN

Niacin, also called *nicotinic acid*, functions in the body as coenzymes in the forms of nicotinamide adenine dinucleotide (NAD) and nicotinamide adenine dinucleotide phosphate (NADP). These coenzymes are hydrogen acceptors; they combine with hydrogen atoms as they are removed from food substrates by many different types of dehydrogenases. Typical operation of both of them is presented in Chapter 67. When a deficiency of niacin exists, the normal rate of dehydrogenation presumably cannot be maintained, and, therefore, oxidative delivery of energy from the foodstuffs to the functioning elements of the cells likewise cannot occur at normal rates.



Niacin

Because NAD and NADP operate in all cells of the body, it is readily understood how lack of niacin can cause multiple symptoms, even though specific metabolic deficiencies have not been pinpointed above other. Clinically, niacin deficiency causes mainly gastrointestinal symptoms, neurologic symptoms, and a characteristic dermatitis. However, it is probably much more proper to say that essentially all functions of the body are depressed or altered.

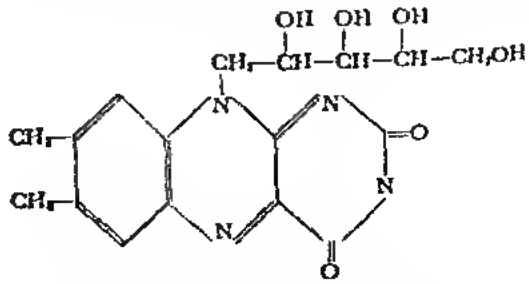
In the early stages of niacin deficiency simple physiological changes such as muscular weakness and poor glandular secretion may occur, but in severe niacin deficiency actual death of tissues ensues. Pathologic lesions appear in many parts of the central nervous system, and permanent dementia or any of many different types of psychoses may result. Also, the skin develops a cracked, pigmented scaliness in areas that are exposed to mechanical irritation or sun irradiation; that it seems as though the skin were unable to repair the different types of irritative damage.

Niacin deficiency causes intense irritation and inflammation of the mucous membranes of the mouth and other portions of the gastrointestinal tract, thus instituting many digestive abnormalities, leading in severe cases to widespread gastrointestinal hemorrhage. It is possible that this results from generalized depression of metabolism in the gastrointestinal epithelium and failure of appropriate epithelial repair.

The clinical entity called *pellagra* and the canine disease called black tongue are caused mainly by niacin deficiency. Pellagra is greatly exacerbated in persons on a corn diet (such as many of the natives of Africa) because corn is very deficient in the amino acid tryptophan, which can be converted in limited quantities to niacin in the body.

RIBOFLAVIN (VITAMIN B_2)

Riboflavin normally combines in the tissues with phosphoric acid to form two coenzymes, *flavin mononucleotide (FMN)*, and *flavin adenine dinucleotide (FAD)*. These in turn operate as hydrogen carriers in the important oxidative systems of the mitochondria. Usually, NAD, operating in association with specific dehydrogenases, accepts hydrogen removed from various food substrates and then passes the hydrogen to FMN or FAD; finally, the hydrogen is released as an ion into the mitochondrial matrix to become oxidized by oxygen, the system for which is described in Chapter 67.



Riboflavin

Deficiency of riboflavin in lower animals causes severe *dermatitis, vomiting, diarrhea, muscular spasticity* which finally becomes muscular weakness, and then *death* preceded by coma and decline in body tempera-

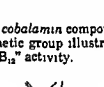
ture. Thus, severe riboflavin deficiency can cause many of the same effects as lack of niacin in the diet; presumably the debilities that result in each instance are due to generally depressed oxidative processes within the cells.

In the human being riboflavin deficiency has never been known to be severe enough to cause the marked debilities noted in animal experiments, but mild riboflavin deficiency is probably common. Such deficiency causes digestive disturbances, burning sensations of the skin and eyes, cracking at the corners of the mouth, headaches, mental depression, forgetfulness, and so on. Perhaps the most common characteristic lesion of riboflavin deficiency is *cheilosis*, which is inflammation and cracking at the angles of the mouth. In addition, a fine, scaly dermatitis often occurs at the angles of the nares, and keratitis of the cornea may occur with invasion of the cornea by small blood vessels.

Though the manifestations of riboflavin deficiency are usually relatively mild, this deficiency frequently occurs in association with deficiency of thiamine or niacin. Therefore, many deficiency syndromes, including *pellagra*, *beriberi*, *sprue*, and *kwashiorkor*, are probably due to a combined deficiency of a number of the vitamins, as well as to other aspects of malnutrition.

VITAMIN B_{12}

Several different *cobalamin* compounds which possess the common prosthetic group illustrated below exhibit so-called "vitamin B_{12}" activity.



Note that this prosthetic group contains cobalt that has coordination bonds similar to those of iron in the hemoglobin molecule. It is likely that the cobalt atom functions in much the same way that the iron atom functions to combine reversibly with other substances.

Vitamin B_{12} performs several metabolic functions, acting as a hydrogen acceptor coenzyme. Its most important function is probably to act as a coenzyme for reducing ribonucleotides to deoxyribonucleotides, a step that is important in the replication of genes. This could explain the two major functions of vitamin B_{12} (1) promotion of growth and (2) promotion of red blood cell formation and maturation. This latter function was described in detail in Chapter 4 in relation to pernicious anemia, a type of anemia caused by failure of red blood cell maturation when vitamin B_{12} is deficient.

A special effect of vitamin B_{12} deficiency is often demyelination of the large nerve fibers of the spinal cord, especially of the posterior columns and occasionally of the lateral columns. As a result, persons with pernicious anemia frequently have much loss of peripheral sensation and, in severe cases, even become paralyzed.

The usual cause of vitamin B_{12} deficiency is not lack of this vitamin in the food but instead deficiency of formation of *intrinsic factor*, which is normally secreted by the parietal cells of the gastric glands and is essential for absorption of vitamin B_{12} by the ileal mucosa. This was discussed in Chapter 4 and in Chapter 66.

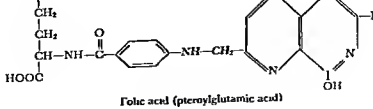
FOLIC ACID (PTEROYLGLUTAMIC ACID)

Several different pteroylglutamic acids, one of which is illustrated below, exhibit the "folic acid effect." Folic acid functions as a carrier of hydroxymethyl and formyl groups. Perhaps its most important use in the body is in the synthesis of purines and thymine, which are required for formation of deoxyribonucleic acid. Therefore, folic acid, like vitamin B_{12}, is also required for replication of the cellular genes. This perhaps explains one of the most important functions of folic acid—that is, to promote growth. Indeed, when it is absent from the diet an animal will grow very little.

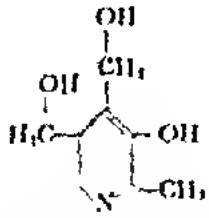
Folic acid is an even more potent growth promoter than vitamin B_{12}, and, like vitamin B_{12}, is also important for the maturation of red blood cells, as discussed in Chapter 4. However, vitamin B_{12} and folic acid each perform specific and different functions in promoting growth and maturation of red blood cells. One of the very significant effects of folic acid deficiency is development of macrocytic anemia. This often can be treated effectively with folic acid alone.

PYRIDOXINE (VITAMIN B_6)

Pyridoxine exists in the form of *pyridoxal phosphate* in the cells and functions as a coenzyme for many different chemical reactions relating to amino acid and protein metabolism. Its most important role is that of coenzyme in the transamination process for the synthesis of amino acids. As a result, pyridoxine plays many key roles in metabolism—especially in protein metabolism. Also, it is believed to act in the transport of some amino acids across cell membranes.



Folic acid (pteroylglutamic acid)

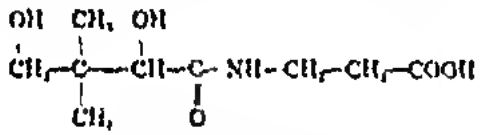


Pyridoxine

Dietary lack of pyridoxine in lower animals can cause dermatitis, decreased rate of growth, development of fatty liver, anemia, and evidence of mental deterioration. Rarely, in children, pyridoxine deficiency has been known to cause convulsions, dermatitis, and gastrointestinal disturbances such as nausea and vomiting.

PANTOTHENIC ACID

Pantothenic acid mainly is incorporated in the body into *coenzyme A*, which has many metabolic roles in the cells. Two of these discussed at length in Chapters 67 and 68 are (1) conversion of decarboxylated pyruvic acid into acetyl-CoA prior to its entry into the citric acid cycle, and (2) degradation of fatty acid molecules into multiple molecules of acetyl CoA. Thus, lack of pantothenic acid can lead to depressed metabolism of both carbohydrates and fats.

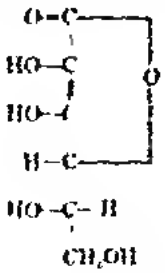


Pantothenic acid

Deficiency of pantothenic acid in lower animals can cause retarded growth, failure of reproduction, graying of the hair, dermatitis, fatty liver, and hemorrhagic adrenal cortical necrosis. In the human being no definite deficiency syndrome has been proved, presumably because of the wide occurrence of this vitamin in almost all foods and because small amounts of the vitamin can probably be synthesized in the body. Nevertheless, this does not mean that pantothenic acid is not of value in the metabolic systems of the body. Indeed, it is perhaps as necessary as any other vitamin.

ASCORBIC ACID (VITAMIN C)

Ascorbic acid is essential for activating the enzyme *prolyl hydroxylase* that promotes the hydroxylation step in the formation of hydroxyproline, an integral constituent of collagen. Without ascorbic acid the collagen that is formed is defective and weak. Therefore, this vitamin is essential for growth of subcutaneous tissue, cartilage, bone, and teeth.



Ascorbic acid (vitamin C)

Deficiency of ascorbic acid for 20 to 30 weeks, as occurred frequently during long sailing voyages in olden days, causes *scurvy*, some effects of which are the following:

One of the most important effects of scurvy is *failure of wounds to heal*. This is caused by failure of the cells to deposit collagen fibrils and intercellular cement substances. As a result, healing of a wound may require several months instead of the several days ordinarily necessary.

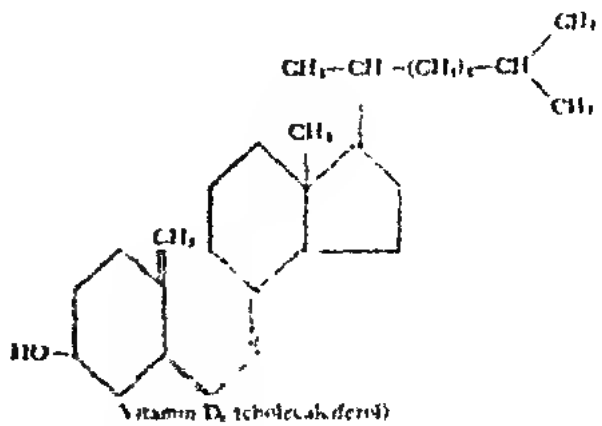
Lack of ascorbic acid causes *cessation of bone growth*. The cells of the growing epiphyses continue to proliferate, but no new collagen is laid down between the cells, and the bones fracture easily at the point of growth because of failure to ossify. Also, when an already ossified bone fractures in a person with ascorbic acid deficiency, the osteoblasts cannot form new bone matrix. Consequently, the fractured bone does not heal.

The *blood vessel walls become extremely fragile* in scurvy, because of failure of the endothelial cells to be cemented together properly and failure to form the collagen fibrils normally present in vessel walls. The capillaries especially are likely to rupture, and as a result many small petechial hemorrhages occur throughout the body. The hemorrhages beneath the skin cause purpuric blotches, sometimes over the entire body. To test for ascorbic acid deficiency, one can produce such petechial hemorrhages by inflating a blood pressure cuff over the upper arm; this occludes the venous return of blood, the capillary pressure rises, and red blotches occur in the skin of the forearm if there is a sufficiently severe ascorbic acid deficiency.

In extreme scurvy the muscle cells sometimes fragment, lesions of the gums with loosening of the teeth occur; infections of the mouth develop; vomiting of blood, bloody stools, and cerebral hemorrhage can all occur and, finally, high fever often develops before death.

VITAMIN D

Vitamin D increases calcium absorption from the gastrointestinal tract and also helps control calcium deposition in the bone. The mechanism by which vitamin D increases calcium absorption is to promote active transport of calcium through the epithelium of the ileum. It especially increases the formation of a calcium-binding protein in the epithelial cells that aids in calcium absorption. The specific functions of vitamin D in relation to overall body calcium metabolism and to bone formation are presented in Chapter 79.



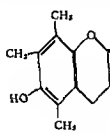
Vitamin D_3 (cholecalciferol)

Several vitamin D compounds exist, one of which is illustrated. This is the natural vitamin D, *cholecalciferol*, which results from ultraviolet irradiation of 7-dehydrocholesterol in the skin. A synthetic vitamin D compound, *ergocalciferol*, is formed by irradiation of ergosterol and is used to a major extent in vitamin D therapy. Both of these vitamin Ds are further altered, first in the liver and then in the kidneys, finally forming *1,25-dihydroxycholecalciferol*, which is the active form of the vitamin. This will be discussed in more detail in Chapter 79.

VITAMIN E

Several related compounds, one of which is illustrated below, exhibit so-called "vitamin E activity." Only rare instances of vitamin E deficiency have occurred in human beings. In lower animals, lack of vitamin E can cause degeneration of the germinal epithelium in the testis and therefore can cause male sterility. Lack of vitamin E can also cause resorption of a fetus after conception in the female. Because of these effects of vitamin E deficiency, vitamin E is sometimes called the "antisterility vitamin."

Vitamin E deficiency in animals can also cause paralysis of the hindquarters, and pathological changes occur in the muscles similar to those found in muscular dystrophy, the disease entity of the human being. However, administration of vitamin E to patients with muscular dystrophy has not proved to be of any benefit.



Vitamin E (alpha-tocopherol)

Finally, as is true of almost all the vitamins, deficiency of vitamin E prevents normal growth, and it sometimes causes degeneration of the renal tubular cells.

Vitamin E is believed to function mainly in relation to unsaturated fatty acids, providing a protective role to prevent oxidation of the unsaturated fats. In the absence of vitamin E, the quantity of unsaturated fats in the cells becomes diminished, causing abnormal structure and function of such cellular organelles as the mitochondria, the lysosomes, and even the cell membrane. Indeed, the muscular dystrophy-like syndrome that occurs in the vitamin E deficiency probably results from continual rupture of lysosomes and subsequent autodigestion of the muscle.

VITAMIN K

Vitamin K is necessary for the formation by the liver of prothrombin, factor VII (proconvertin), factor IX, and factor X, all of which are important in blood coagulation. Therefore, when vitamin K deficiency occurs, blood clotting is retarded. The function of this vitamin and its relationships with some of the anticoagulants, such as dicumarol, have been presented in greater detail in Chapter 8.

Several different compounds, both natural and synthetic, exhibit vitamin K activity. The chemical formula for one of the natural vitamin K compounds is illustrated below. Because vitamin K is synthesized by bacteria in the colon, a dietary source for this vitamin is not usually necessary, but, when the bacteria of the colon are destroyed by administration of large quantities of antibiotic drugs, vitamin K deficiency occurs rapidly because of the paucity of this compound in the normal diet.

MINERAL METABOLISM

The functions of many of the minerals, such as sodium, potassium, chloride, and so forth, have been presented at appropriate points in the text. Therefore, only specific functions of minerals not covered elsewhere are mentioned here.

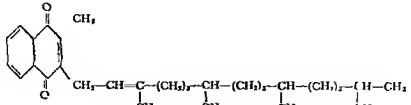
The body content of the most important minerals is listed in Table 73–3, and the daily requirements of these are given in Table 73–4.

Magnesium. Magnesium is approximately one sixth as plentiful in cells as potassium. Magnesium is especially required as a catalyst for many intracellular enzymatic reactions, particularly those relating to carbohydrate metabolism.

The extracellular magnesium concentration is slight, only 1.8 to 2.5 mEq/liter. Increased extracellular concentration of magnesium depresses activity in the ner-

Table 73–3 CONTENT IN GRAMS OF A 70 KILOGRAM ADULT MAN

Water	41,400	Mg	21
Fat	12,600	Cl	85
Protein	12,600	P	670
Carbohydrate	300	S	112
Na	63	Fe	3
K	150	I	0.014
Ca	1,160		



Vitamin K₁ (2-methyl-3-phytyl-1,4-naphthoquinone)

DiSilvestro, R A, and Cousins, R J Physiological ligands for copper and zinc *Annu Rev Nutr*, 3:261, 1983.
Fehg, P: Starvation. *In* DeGroot, L J, *et al* (eds.) Endocrinology. Vol 3. New York, Grune & Stratton, 1979, p 1927
Festing, M. F. W. (ed). Animal Models of Obesity New York, Oxford University Press, 1979.
Forbes, R. M, and Erdman, J W, Jr Bioavailability of trace mineral elements *Annu Rev. Nutr*, 3 213, 1983
Gershwin, M E Nutrition and Immunity New York, Academic Press, 1985.
Guengerich, F. P Effects of nutritive factors on metabolic processes involving bioactivation and detoxication of chemicals *Annu. Rev Nutr*, 4 207, 1984
Havel, R. J Calorie homeostasis and disorders of fuel transport. *N Engl. J Med.*, *287* 1186, 1972
Henderson, L. M Niacin *Annu Rev Nutr*, 3 289, 1983
Hoebel, B, G Feeding: Neural control of intake *Annu Rev Physiol*, 33 533, 1971
Howard, L., and Michalek, A. V., Home parenteral nutrition (HPN) *Annu Rev Nutr*, 4 69, 1984
Hunt, S M, *et al* Nutrition. Principles and Clinical Practice New York, John Wiley & Sons, 1980
Hurley, L. S, Teratogenic aspects of manganese, zinc, and copper nutrition *Physiol Rev*, *61* 249, 1981
Karcioglu, Z. A, and Sarper, R M (eds) Zinc and Copper in Medicine Springfield, Ill, Charles C Thomas, 1980
Keesey, R. E Neurophysiologic control of body fatness *In* Laws, R M, and Shekelle, R B (eds) Childhood Prevention of Atherosclerosis and Hypertension New York, Raven Press, 1980
LeBow, M. D Weight Control The Behavioural Strategies New York, John Wiley & Sons, 1980
MacLean: Pediatric Nutrition in Clinical Practice Menlo Park, Addison-Wesley Publishing Co, 1984
Magnen, J. L. Body energy balance and food intake A neuroendocrine regulatory mechanism *Physiol Rev*, 63 314, 1983
Marshall, C W Vitamins and Minerals, Help or Harm? Philadelphia, G F Stickley, 1983
Merr, M L Introductory Foods New York, McMillan Publishing Co, 1985
Nielsen, F H. Ultratrace elements in nutrition *Annu Rev Nutr*, 4 21, 1984
Olson, R. E. The function and metabolism of vitamin K *Annu. Rev Nutr*, 4:281, 1984
Oomura, Y, and Yoshimatsu, H. Neural network of glucose monitoring system *J Auton Nerv Syst*, *10* 359, 1984
Peng, S K., and Taylor, C B Cholesterol autoxidation, health and atherosclerosis A review of situations in developed countries. *World Rev Nutr Diet*, *44* 117, 1984
Rude, R. K., and Singer, F R Magnesium deficiency and excess *Annu Rev Med*, 32 245, 1981
Ryan, U S Structural bases for metabolic activity *Annu Rev Physiol*, *44* 223, 1982
Salans, L. B Obesity and the adipose cell *In* Bondy, P K, and Rosenberg, L. E. (eds): Metabolic Control and Disease 8th Ed Philadelphia, W B Saunders Co., 1980, p 495
Shoden, R. J, and Griffin, W S Fundamentals of Clinical Nutrition New York, McGraw-Hill, 1980
Storlien, L. H The role of the ventromedial hypothalamic area in perprandial glucoregulation *Life Sci.*, *360* 505, 1985
Sutor, C W, and Hunter, M F Principles and Application in Health Promotion Philadelphia, J B Lippincott Co, 1980
Sullivan, A C, and Gruen, R K Mechanisms of appetite modulation by drugs *Fed Proc*, *44* 139, 1985
Swan, P B Food consumption by individuals in the United States. Two major surveys *Annu Rev Nutr*, 3 413, 1983
Thompson, C I Controls of Eating New York, Spectrum Publications, 1979
Tielsch, J M, and Sommer, A The epidemiology of vitamin A deficiency and xerophthalmia *Annu. Rev Nutr*, 4 153, 1984
Waterlow J C (ed) Nitrogen Metabolism in Man New York, Elsevier Science Publishing Co, 1982
Weir, D G, *et al* Folate metabolism, the enterohepatic circulation and alcohol *Biochem Pharmacol*, *34* 1, 1985
Willard. Nutrition for the Practicing Physician Menlo Park, Addison-Wesley Publishing Co., 1984
Wolf, G Multiple functions of vitamin A *Physiol Rev* *64* 873, 1984

XIII

ENDOCRINOLOGY AND REPRODUCTION

73 ■ Dietary Balances; Regulation of Feeding; Obesity and Starvation; Vitamins and Minerals

74 ■ Introduction to Endocrinology

75 ■ The Pituitary Hormones and Their Control by the Hypothalamus

76 ■ The Thyroid Metabolic Hormones

77 ■ The Adrenocortical Hormones

78 ■ Insulin, Glucagon, and Diabetes Mellitus

79 ■ Parathyroid Hormone, Calcitonin, Calcium and Phosphate Metabolism, Vitamin D, Bone, and Teeth

80 ■ Reproductive and Hormonal Functions of the Male

81 ■ Female Physiology Before Pregnancy, and the Female Hormones

82 ■ Pregnancy and Lactation

83 ■ Special Features of Fetal and Neonatal Physiology

74

Introduction to Endocrinology

The functions of the body are regulated by two major control systems: (1) the nervous system, which has been discussed, and (2) the hormonal, or endocrine, system. In general, the hormonal system is concerned principally with control of the different metabolic functions of the body, such as controlling the rates of chemical reactions in the cells or the transport of substances through cell membranes or other aspects of cellular metabolism like growth and secretion. Some hormonal effects occur in seconds, whereas others require several days simply to start and then continue for weeks, months, or even years

Many interrelationships exist between the hormonal and nervous systems. For instance, at least two glands secrete their hormones only in response to appropriate neural stimuli, the *adrenal medullae* and the *pituitary gland* In turn, many of the pituitary hormones control secretion by the majority of the other endocrine glands, as we shall see in the following chapters

NATURE OF A HORMONE

A hormone is a chemical substance that is secreted into the body fluids by one cell or a group of cells and that exerts a physiological *control* effect on other cells of the body

At many points in this text we have already discussed different hormones Some are *local hormones* and others are *general hormones*. Examples of local hormones are acetylcholine released at the parasympathetic and skeletal nerve endings, *secretin* released by the duodenal wall and transported in the blood to the pancreas to cause a watery pancreatic secretion, *cholecystokinin* released in the small intestine and transported to the gallbladder to cause it to contract and to the pancreas to cause enzyme secretion, and many others. These hormones obviously have specific local effects, whence comes the name local hormones

Two examples of general hormones with which we are already familiar are *epinephrine* and *norepinephrine*, both of which are secreted by the adrenal medullae in response to sympathetic stimulation. These hormones are transported in the blood to all parts of the body and cause many different reactions, especially constriction of the blood vessels and elevation of the arterial pressure. Another general hormone is *renin*, which is secreted by the kidneys in response to hypotension; the substance angiotensin II, which is formed in the blood as a result of the enzymatic action of renin on a plasma protein, is also a general hormone.

Most of the general hormones, however, are secreted by *endocrine glands* and are then transported in the blood throughout the body. A few of the general hormones affect all or almost all cells of the body; examples are *growth hormone* from the anterior pituitary gland that causes growth in all or most parts of the body and *thyroid hormone* from the thyroid gland that increases the rates of most chemical reactions in almost all the body's cells.

Other hormones, however, affect only specific tissues, called *target tissues* because only these tissues have the specific *receptors* that will bind the respective hormones to initiate their actions. For instance, adrenocorticotropin from the anterior pituitary gland specifically stimulates the adrenal cortex, causing it to secrete the adrenocortical hormones; and the ovarian hormones have specific effects on the female sex organs, as well as on the secondary sexual characteristics of the female body. Many examples of target tissues will become apparent in the following chapters.

AN OVERVIEW OF THE IMPORTANT ENDOCRINE GLANDS AND THEIR HORMONES

Figure 74–1 illustrates the anatomical loci of the important endocrine glands of the body except for the testes and exrept for the placenta, which is an additional important source of sex hormones during pregnancy. Let us give a preview of the important hormones secreted by these glands and their most important actions:

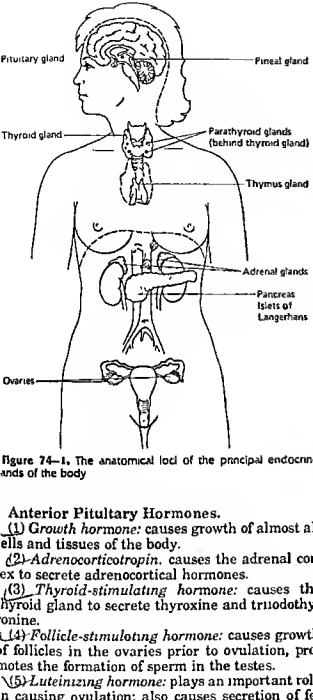


Figure 74–1. The anatomical loci of the principal endocrine glands of the body

Anterior Pituitary Hormones.

(1) *Growth hormone:* causes growth of almost all cells and tissues of the body.

(2) *Adrenocorticotropin.* causes the adrenal cortex to secrete adrenocortical hormones.

(3) *Thyroid-stimulating hormone:* causes the thyroid gland to secrete thyroxine and triiodothyronine.

(4) *Follicle-stimulating hormone:* causes growth of follicles in the ovaries prior to ovulation, promotes the formation of sperm in the testes.

(5) *Luteinizing hormone:* plays an important role in causing ovulation; also causes secretion of female sex hormones by the ovaries and testosterone by the testes.

(6) *Prolactin:* promotes development of the breasts and secretion of milk.

Posterior Pituitary Hormones.

(1) *Antidiuretic hormone* (also called *vasopressin*): causes the kidneys to retain water, thus increasing the water content of the body; also, in high concentrations, causes constriction of the blood vessels throughout the body and elevates the blood pressure.

(2) *Oxytocin.* contracts the uterus during the birthing process, thus perhaps helping expel the baby; also contracts myoepithelial cells in the breasts, thereby expressing milk from the breasts when the baby suckles.

Adrenal Cortex.

(1) *Cortisol:* has multiple metabolic functions for control of the metabolism of proteins, carbohydrates, and fats.

(2) *Aldosterone:* reduces sodium excretion by the kidneys and increases potassium excretion, thus increasing sodium in the body while decreasing the amount of potassium.

Thyroid Gland.

(1 and 2) *Thyroxine* and *triiodothyronine:* increase the rates of chemical reactions in almost all cells of the body, thus increasing the general level of body metabolism.

(3) *Calcitonin:* promotes the deposition of calcium in the bones and thereby decreases calcium concentration in the extracellular fluid.

Islets of Langerhans in the Pancreas.

(1) *Insulin:* promotes glucose entry into most cells of the body, in this way controlling the rate of metabolism of most carbohydrates.

(2) *Glucagon:* increases the release of glucose from the liver into the circulating body fluids.

Ovaries.

(1) *Estrogens:* stimulate the development of the female sex organs, the breasts, and various secondary sexual characteristics.

(2) *Progesterone:* stimulates secretion of "uterine milk" by the uterine endometrial glands; also helps promote development of the secretory apparatus of the breasts.

Testes.

(1) *Testosterone:* stimulates growth of the male sex organs; also promotes the development of male secondary sex characteristics.

Parathyroid Gland.

(1) *Parathormone:* controls the calcium ion concentration in the extracellular fluid by controlling (a) absorption of calcium from the gut, (b) excretion of calcium by the kidneys, and (c) release of calcium from the bones

Placenta.

(1) *Human chorionic gonadotropin:* promotes growth of the corpus luteum and secretion of estrogens and progesterone by the corpus luteum.

(2) *Estrogens:* promote growth of the mother's sex organs and of some of the tissues of the fetus.

(3) *Progesterone:* probably promotes development of some of the fetal tissues and organs; helps promote development of the secretory apparatus of the mother's breasts.

(4) *Human somatomammotropin:* probably promotes growth of some fetal tissues as well as aiding in the development of the mother's breasts.

From this overview of the endocrine system, it is clear that most of the metabolic functions of the body are controlled one way or another by the endocrine glands. For instance, without growth hormone, the person remains a dwarf. Without thyroxine and triiodothyronine from the thyroid gland, almost all the chemical reactions of the body become sluggish, and the person becomes

sluggish as well. Without insulin from the pancreas, the body's cells can utilize very little of the food carbohydrates for energy. And, without the sex hormones, sexual development and sexual functions are absent.

CHEMISTRY OF THE HORMONES

Chemically, the hormones are of three basic types

(1) *Steroid hormones:* These all have a chemical structure similar to that of cholesterol and in most instances are derived from cholesterol itself. Different steroid hormones are secreted by (a) the adrenal cortex (*cortisol* and *aldosterone*), (b) the ovaries (*estrogen* and *progesterone*), (c) the testes (*testosterone*), and (d) the placenta (*estrogen* and *progesterone*).

(2) *Derivatives of the amino acid tyrosine:* Two groups of hormones are derivatives of the amino acid tyrosine The two metabolic thyroid hormones, *thyroxine* and *triiodothyronine*, are iodinated forms of tyrosine derivatives. And the two principal hormones of the adrenal medullae, *epinephrine* and *norepinephrine*, are both catecholamines, also derived from tyrosine.

(3) *Proteins or peptides:* All the remaining important endocrine hormones are either proteins, peptides, or immediate derivatives of these The anterior pituitary hormones are either proteins or large polypeptides, the posterior pituitary hormones, antidiuretic hormone and oxytocin, are peptides containing only eight amino acids And insulin, glucagon, and parathormone are all large polypeptides

Storage and Secretion of Hormones

We shall see in the following chapters that there is no single way in which all the endocrine glands store and secrete their hormones However, several general patterns occur for most of the hormones For instance, all the protein hormones are formed by the granular endoplasmic reticulum of the glandular cells in the same manner that other secretory proteins are formed as was described in Chapter 3 However, the initial protein formed by the endoplasmic reticulum almost never is the final hormone itself Instead, it is larger than the active hormone and is called a *preprohormone* Then, this large protein is further cleaved, usually while still in the endoplasmic reticulum, to form a smaller protein called the *prohormone*. This in turn is transported in the ER transport vesicles to the Golgi apparatus where still another section of the protein is cleaved, in this way the final active protein hormone is formed The Golgi apparatus usually also compacts the hormone molecules into small membrane-encapsulated vesicles called *secretory vesicles* or *secretory granules* These then remain stored in the cytoplasmic compartment of [illegible] cell until a specific signal, such as a nerve signal, another hormonal signal, or a local chemical or physical signal, comes along to cause secretion.

The two groups of hormones derived from tyrosine, the thyroid and adrenal medullary hormones, are both formed by the actions of enzymes in the cytoplasmic compartments of the glandular cells. In the case of the adrenal medullary hormones norepinephrine and epinephrine, these are then absorbed into preformed vesicles and stored until they are to be secreted. The thyroid metabolic hormones, thyroxine and triiodothyronine, on the other hand, are formed as component parts of a large protein molecule called thyroglobulin, and this is then stored in large follicles within the thyroid gland When the thyroid hormones are to be secreted, specific enzyme systems within the thyroid glandular cells cleave the thyroglobulin molecule, thereby allowing the thyroid hormones to be released into the blood.

For the steroid hormones secreted by the adrenal cortex, the ovaries, or the testes, the amounts stored in the glandular cells are usually quite small, but large amounts of precursor molecules, especially cholesterol and various intermediates between cholesterol and the final hormones, are present in the cells. Upon appropriate stimulation, enzymes in these cells can within minutes cause the chemical conversions to the final hormones, followed almost immediately by secretion.

Onset of Hormone Secretion Following a Stimulus, and Durations of Action of Different Hormones. Some hormones are secreted within seconds, after the gland is stimulated, and they may develop full action within another few seconds to minutes; the actions of other hormones often require months for full effect. To give examples, the two adrenal medullary hormones, epinephrine and norepinephrine, begin to be secreted in response to sympathetic nerve stimuli within the first second of stimulation and usually reach maximum activity within a minute after the onset of stimulation However, these hormones are also destroyed rapidly by local tissue enzymes, or they are absorbed into cells Therefore, their duration of action is usually no more than one to three minutes at most after the sympathetic stimulation is over

At the other extreme, the thyroid hormones are stored in the form of thyroglobulin in the thyroid follicles, sometimes for several months, before finally being secreted Then, after secretion, several hours to several days are required before even initial activity occurs; but their effect on enhancement of tissue metabolism can last as long as four to six weeks Thus, each of the different hormones has its own characteristic onset and duration of action—each tailored to perform its specific control function

Concentrations of Hormones in the Circulating Blood, and Hormonal Secretion Rates. The quantitative amounts of hormones that are

required to control most metabolic and endocrine functions are incredibly small. Their concentrations in the blood range from as little as 1 picogram (which is 1 millionth of a millioth of a gram) in each milliliter of blood up to at most a few micrograms (1 millionth of a gram) per milliliter of blood. Similarly, the rates of secretion of the various hormones are also extremely small, measured in micrograms or milligrams per day. We shall see later in this chapter that highly specialized mechanisms are available in the target tissues that allow even these minute quantities of hormones to exert powerful control over the physiologic systems.

Control of Hormone Secretion Rate—The Role of Negative Feedback

Without exception, the rate of secretion of every hormone that has ever been studied is itself controlled very exactly by some internal control system. In most instances, this control is exerted through a negative feedback mechanism as follows:

(1) The endocrine gland has a natural tendency to *oversecrete* its hormone.

(2) Because of this tendency, the hormone exerts more and more of its control effect on the target organ.

(3) The target organ in turn performs its function.

(4) But when too much function occurs, usually some factor about the function then *feeds back* to the endocrine gland and causes a *negative* effect on the gland to decrease its secretory rate. Thus, the function of the hormone is monitored, and this information in turn provides negative feedback control of the secretory rate by the gland.

Careful consideration of the above negative feedback mechanism will show that the important factor to be controlled usually is not the secretory rate of the hormone but instead the degree of activity of the target organ. Therefore, only when the target organ activity rises to an appropriate level will the feedback to the gland become powerful enough to slow further secretion of the hormone. If the target organ responds poorly to the hormone, the endocrine gland will almost always secrete still more and more of its hormone until the target organ eventually reaches the appropriate level of activity, but at the expense of excessive secretion of the controlling hormone.

HORMONE RECEPTORS AND THEIR ACTIVATION

The endocrine hormones almost never act directly on the intracellular machinery to control the different cellular chemical reactions; instead, they almost invariably first combine with *hormone receptors* on the surfaces of the cells or inside the cells. The combination of hormone and receptor then usually initiates a cascade of reactions in the cell.

Either all or almost all hormonal receptors are very large proteins, and each cell usually has some 2000 to 10,000 receptors.

Also, each receptor is usually highly specific for a single hormone; this determines the type of hormone that will act on a particular tissue. Obviously, the target tissues that are affected by a hormone are those that contain its specific receptors.

The locations of the receptors for the different types of hormones are generally the following:

(1) *In the membrane.* The membrane receptors are specific mostly to the protein, peptide, and catecholamine (epinephrine and norepinephrine) hormones.

(2) *In the cytoplasm.* The receptors for the different steroid hormones are found almost entirely in the cytoplasm.

(3) *In the nucleus.* The receptors for the metabolic thyroid hormones (thyroxine and triiodothyronine) are found in the nucleus, believed to be located in direct association with one or more of the chromosomes.

Regulation of the Number of Receptors. The number of receptors in a target cell often does not remain constant from day to day, or even from minute to minute, for the receptor proteins are themselves often destroyed or new ones are manufactured by the protein-manufacturing mechanism of the cell. For instance, binding of a hormone with its target cell receptors usually causes the number of receptors to decrease, either because of inactivation of some of the receptor molecules or because of decreased production of the molecules. In either event, this is called *down-regulation* of the receptors. Obviously, this decreases the responsiveness of the target tissue to the hormone over a period of time as the number of receptors decreases.

In a few instances, hormones cause *up-regulation* of receptors; that is, the stimulating hormone induces the formation of more receptor molecules by the protein-manufacturing machinery of the target cell. In this instance, the target tissue becomes progressively more sensitive to the stimulating effects of the hormone.

MECHANISMS OF HORMONAL ACTION

Activation of the Receptors. The receptors in their unbound state usually are inactive, and the intracellular mechanisms that are associated with them are also inactive. However, in a few instances the unbound receptors are in the active form, and when bound with the hormone they become inhibited.

Activation of a receptor occurs in different ways for different types of receptors. For instance, in

Chapter 46 we discussed the many types of receptors located in the postsynaptic membranes of neurons and activated by the synaptic hormones called "transmitter substances." In general, the transmitter substance combines with the receptor and causes a conformational change of the receptor molecule; this in turn alters the membrane permeability to one or more ions, especially sodium, chloride, potassium, and calcium ions. A few of the general endocrine hormones also function in this same way—for instance, the effect of epinephrine and norepinephrine in changing the membrane permeability in certain of their target tissues.

In addition to this occasional direct effect of hormone receptors to change cell membrane permeability, there are also two very important general mechanisms by which a large share of the hormones function: (1) by *activating the cyclic AMP system of the cells*, which in turn activates multiple other intracellular functions; or (2) by *activating the genes of the cell*, which cause the formation of intracellular proteins that in turn initiate specific cellular functions. These two general mechanisms are described as follows:

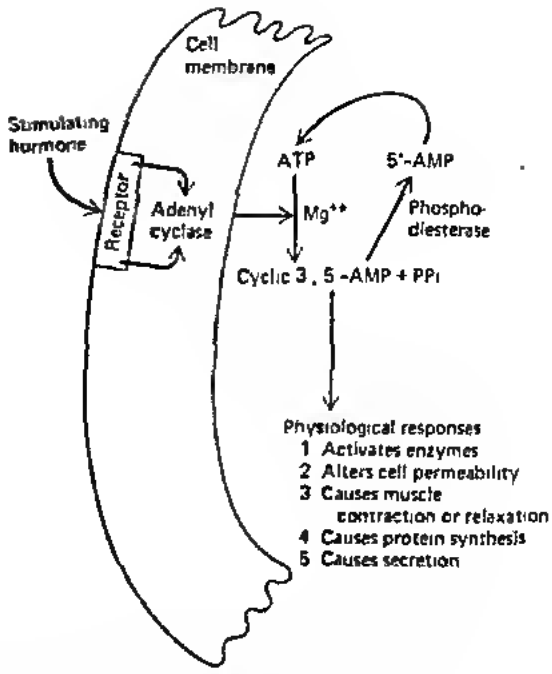


Figure 74–2. The cyclic AMP mechanism by which many hormones exert their control of cell function

THE CYCLIC AMP MECHANISM FOR CONTROLLING CELL FUNCTION—A "SECOND MESSENGER" FOR HORMONE MEDIATION

Many hormones exert their effects on cells by first causing the substance *cyclic 3',5'-adenosine monophosphate* (cyclic AMP) to be formed in the cell. Once formed, the cyclic AMP causes the hormonal effects inside the cell. Thus, *cyclic AMP is an intracellular hormonal mediator.* It is also frequently called a *second messenger* for hormone mediation—the "first messenger" being the original stimulating hormone.

The cyclic AMP mechanism has been shown to be a way in which all the following hormones (and many more) can stimulate their target tissues:

1. Adrenocorticotropin
2. Thyroid-stimulating hormone
3. Luteinizing hormone
4. Follicle-stimulating hormone
5. Vasopressin
6. Parathyroid hormone
7. Glucagon
8. Catecholamines
9. Secretin
10. The hypothalamic releasing hormones.

Figure 74–2 illustrates the function of the cyclic AMP mechanism in more detail. The stimulating hormone first binds with a specific "receptor" for that hormone on the membrane surface of the target cell. The specificity of the receptor determines which hormone will affect the target cell. After binding with the membrane receptor, the combination of hormone and receptor activates the protein enzyme *adenyl cyclase.* This enzyme is also located in the membrane and is either bound directly with the receptor protein or closely associated with it. However, a large portion of the adenyl cyclase enzyme protrudes through the inner surface of the membrane into the cytoplasm and, when activated, causes immediate *conversion of much of the cytoplasmic ATP into cyclic AMP.*

Once cyclic AMP is formed inside the cell it activates still other enzymes. In fact, it usually activates a *cascade of enzymes.* That is, a first enzyme is activated and this activates another enzyme, which activates still a third, and so forth. The importance of this mechanism is that only a few molecules of activated adenyl cyclase in the cell membrane can cause many more molecules of the next enzyme to be activated, which can cause still many times that many molecules of the third enzyme to be activated, and so forth. In this way, even the slightest amount of hormone acting on the cell surface can initiate a very powerful cascading activating force for the entire cell.

The specific action that occurs in response to cyclic AMP in each type of target cell depends upon the nature of the intracellular machinery, some cells having one set of enzymes and other cells having other enzymes. Therefore, different functions are elicited in different target cells—such functions as

(1) initiating synthesis of specific intracellular chemicals,
(2) causing muscle contraction or relaxation,
(3) initiating secretion by the cells,
(4) altering the cell permeability,
(5) and many other possible effects.

Thus, a thyroid cell stimulated by cyclic AMP forms the metabolic hormones thyroxine and triiodothyronine, whereas the same cyclic AMP in an adrenocortical cell causes secretion of the adrenocortical steroid hormones. On the other hand, cyclic AMP affects epithelial cells of the renal tubules by increasing their permeability to water.

Role of Calcium Ions and "Calmodulin" as a Second Messenger System. Another second messenger system operates in response to the entry of calcium ions into cells The calcium entry may be initiated by electrical phenomena that open membrane calcium channels or by hormones interacting with membrane receptors that similarly open calcium channels.

On entering the cell, the calcium ions bind with a protein called *calmodulin* This protein has four separate calcium sites; when as many as three or four of these sites have bound with calcium, a conformational change occurs that activates the calmodulin, causing multiple effects inside the cell in the same way that cyclic AMP functions For instance, it activates many other enzymes in addition to those activated by cyclic AMP, thus causing an additional set of intracellular metabolic reactions. One of the specific functions of calmodulin is to activate myosin kinase that then acts directly on the myosin of smooth muscle to cause smooth muscle contraction.

The normal calcium ion concentration in most cells of the body is about 10^{-7} to 10^{-8} millimoles per liter, which is not enough to activate the calmodulin system. But, when the calcium ion concentration rises to as high as 10^{-6} to 10^{-5} millimoles per liter, enough binding then occurs to cause all the intracellular actions of calmodulin. This is almost exactly the same amount of calcium ion change that is required in skeletal muscle to activate troponin C, which in turn causes skeletal muscle contraction. It is interesting that troponin C is very similar to calmodulin both in function and in protein structure.

Other Intracellular Hormonal Mediators. Still other intracellular mediators (or "second messengers") also exist in the cell. Two of these are (1) *cyclic guanosine monophosphate*, which is nucleotide that functions almost exactly like cyclic AMP except that it contains the base guanine rather than adenine and it activates a different set of enzymes; and (2) *inositol triphosphate*, which can also be formed in response to hormone-binding with specific cell membrane receptors—and this substance, too, activates still other enzymes.

Finally, it is probable that many *prostaglandins* also function as second messenger substances. These are a series of lipid compounds closely related to each other and widely present in cells throughout the body. Hundreds of different cellular control functions have been postulated for the different prostaglandins, though most of these functions still await definitive proof.

The importance of the multiple second messenger systems within the cells is that each of them can control separate cellular functions, thus allowing multiple modes of cellular action.

ACTION OF STEROID HORMONES ON THE GENES TO CAUSE PROTEIN SYNTHESIS

A second major means by which hormones—specifically the steroid hormones secreted by the adrenal cortex, the ovaries, and the testes—act is to cause synthesis of proteins in the target cells; these proteins then function as enzymes or transport proteins that in turn activate other functions of the cells.

The sequence of events in steroid function is the following:

1. The steroid hormone enters the cytoplasm of the cell, where it binds with a specific *receptor protein*.
2. The combined receptor protein/hormone then diffuses into or is transported into the nucleus.
3. The combination now activates the transcription process of specific genes to form messenger RNA.
4. Tha messenger RNA diffuses into the cytoplasm where it promotes tha translation process at the ribosomes to form new proteins.

To give an example, aldosterone, one of the hormones secreted by the adrenal cortex, enters the cytoplasm of renal tubular cells, which contain its specific receptor protein. Therefore, in these cells the above sequence of avents ensues. After about 45 minutes, proteins begin to appear in the ranal tubular cells that promote sodium reabsorption from tha tubules and potassium secretion into the tubules. Thus, there is a characteristic delay in the beginning action of the steroid hormona of 45 minutes and up to several hours or even days for full action, which is in marked contrast to the almost instantaneous action of some of the peptide and amino acid-derived hormones, such as vasopressin and norepinephrine.

ACTION OF THE THYROID HORMONES IN THE CELL NUCLEUS

The thyroid hormones thyroxine and triiodothyronine cause increased transcription by certain genes in the nucleus. To accomplish this, these hormones first bind directly with receptor proteins in the nucleus itself; these receptors are probably protein molecules located within the chromosomal complex, and they probably control the function of the genetic promotors or operators as explained in Chapter 3.

Two important features of thyroid hormone function in the nucleus are:

(1) They activate the genetic mechanisms for formation of many different types of intracellular proteins—probably a hundred or more. Many of

these are enzymes that promote enhanced intracellular metabolic activity.

(2) Once bound to the intranuclear receptors, the thyroid hormones can continue to express their control functions for days or even weeks.

MEASUREMENT OF HORMONE CONCENTRATIONS IN THE BLOOD

Most hormones are present in the blood in extremely minute quantities, some in concentrations as low as one millionth of a milligram (1 picogram) per milliliter. Therefore, except in a few instances, it has been almost impossible to measure these concentrations by usual chemical means. Fortunately, though, an extremely sensitive method was developed about two decades ago that revolutionized the measurement of hormones, their precursors, and their metabolic end-products. This is the method of *radioimmunoassay*.

Radioimmunoassay

The principle of radioimmunoassay is the following:

First, an antibody is developed that is highly specific for the hormone to be measured.

Second, a small quantity of this antibody is mixed with (a) a quantity of fluid from the animal containing the hormone to be measured and (b) an appropriate amount of purified standard hormone of the same type that has been tagged with a radioactive isotope. However, one specific condition must be met: there must be too little antibody to bind completely both the tagged hormone and the hormone in the fluid to be assayed. Therefore, the natural hormone in the assay fluid and the radioactive standard hormone *compete for the binding sites* on the antibody. In the process of competing, the quantity of each of the two hormones, the natural and the radioactive, that binds is proportional to its concentration.

Third, after binding has reached equilibrium, the antibody-hormone complex is separated from the remainder of the solution, and the quantity of radioactive hormone bound with antibody is measured by radioactive counting techniques. If a *large amount of radioactive hormone* has bound with the antibody, then it is clear that there was only a *small amount of natural hormone* to compete with the radioactive hormone, and therefore the concentration of the natural hormone in the assayed fluid was small. Conversely, if only a small amount of radioactive hormone has bound, it is clear that there was a very large amount of natural hormone to compete for the binding sites.

Fourth, to make the assay highly quantitative, the radioimmunoassay procedure is performed also for "standard" solutions of untagged hormone at several different concentration levels. Then, a "standard curve" is plotted as illustrated in Figure 74–3. By comparing the radioactive counts recorded from the original assay procedure with the standard curve, one can determine within an error of ± 10 to 15 per cent the concentration of the hormone in the assayed fluid. As little as one trillionth of a gram of hormone is often assayed in this way.

Other Competitive Binding Assay Procedures. Other competitive binding techniques for assay of minute quantities of hormones have also been employed. One of these is to use in place of the antibody a specific carrier globulin of the plasma that is a natural binding agent for some specific hormone. For instance, the plasma-binding globulin for the adrenocortical hormone cortisol is highly specific. Therefore, this globulin is substituted for the antibody in the assay process for cortisol, the assay is then carried out in exactly the same way as the radioimmunoassay procedure.

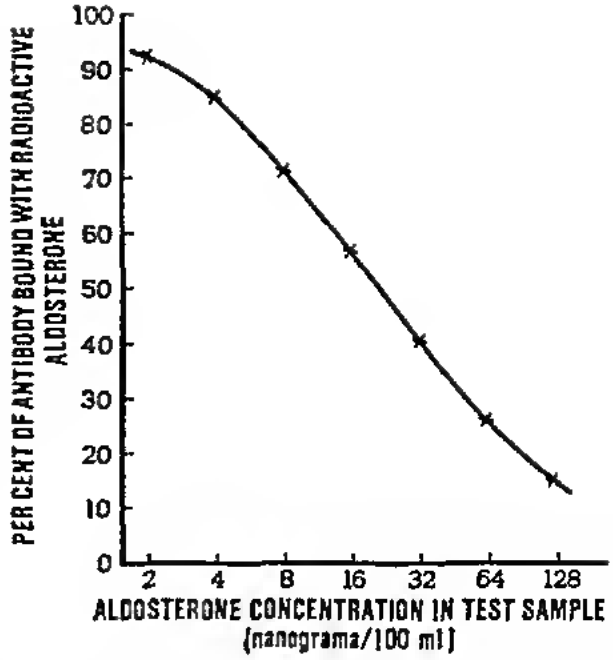


Figure 74–3. A "standard curve" for radioimmunoassay of aldosterone (Courtesy of Dr. Manis Smith.)

Metabolic Clearance Rate of Hormones

Two different factors can increase or decrease the concentration of a hormone in the blood. One of these, obviously, is the rate of secretion of the hormone. The second is the rate of removal of the hormone from the blood, which is called the *metabolic clearance rate*. This is expressed in terms of the number of milliliters of plasma cleared of the hormone per minute. To calculate this clearance rate, one makes the following two measurements:

(1) The rate of disappearance of the hormone from the plasma per minute.

(2) The concentration of the hormone in each milliliter of plasma.

Then, the metabolic clearance rate is calculated by the following formula:

$$\text{Metabolic clearance rate} = \frac{\left\{\text{Rate of disappearance of hormone from the plasma}\right\}}{\left\{\text{Concentration of hormone in each milliliter of plasma}\right\}}$$

The usual procedure for making this measurement is the following: A purified solution of the hormone to be measured is tagged with a radioactive substance. Then the radioactive hormone is infused at a constant rate into the blood stream until the radioactive concentration in the plasma becomes steady. At this time, the rate of disappearance of the radioactive hormone from the

plasma obviously equals the rate at which it is infused, which gives one the rate of disappearance. At the same time, the plasma concentration of the radioactive hormone is measured using a standard radioactive counting procedure. Then, using the formula just cited, the metabolic clearance rate is calculated.

Hormones are "cleared" from the plasma in several ways, including (1) metabolic destruction by the tissues, (2) binding with the tissues, (3) excretion by the liver into the bile, and (4) excretion by the kidneys into the urine. For certain hormones, a decreased metabolic clearance rate frequently causes an excessively high concentration of the hormone in the circulating body fluids. For instance, this occurs for several of the steroid hormones when the liver is diseased because these hormones are mainly conjugated in the liver and then "cleared" into the bile

Rate of Hormone Secretion

A simple method for estimating the rate of hormone secretion is, first, to measure the concentration of the natural hormone in the plasma using a radioimmunoassay procedure Next, the metabolic clearance rate is measured. Then, by multiplying the concentration of the natural hormone times the metabolic clearance rate, one derives a value that is equal to the steady-state rate of hormone production

However, hormone production often increases and decreases very rapidly In such cases, one can measure the changing rates of secretion only by collecting samples of both the arterial blood entering the gland and the venous blood leaving the gland, and simultaneously measuring the rate of blood flow through the gland. Then, by multiplying the rate of blood flow times the venous increase in hormone concentration over that in the arterial blood, one can derive the instantaneous secretion rate Using methods such as this, it has been possible to show that many endocrine glands can be stimulated to extremely high secretory rates within minutes, and many glands also secrete hormones very intermittently rather than steadily This is especially true for the pituitary hormones as well as for cortisol from the adrenal gland.

REFERENCES

Almers, W , and Stirling, C Distribution of transport proteins over animal cell membranes *J Membr Biol*, *77*169, 1984

Aurbach, G D Polypeptide and amine hormone regulation of adenylate cyclase. *Annu Rev Physiol 44* 653, 1982

Blackshear, P J Key References in Endocrinology An Annotated Guide New York, Churchill Livingstone, 1982

Charlton, H M. Mouse mutants as models in endocrine research. *Q. J Exp Physiol*, *69* 655, 1984

Cheung, W Y Calmodulin Its potential role in cell proliferation and heavy metal toxicity *Fed Proc*, *43* 2995, 1984

Daly, J. W Forskolin, adenylate cyclase, and cell physiology. an overview, *Adv Cyclic Nucleotide Res*, *17* 81, 1984

Dawson, K G · Endocrine physiology of electrolyte metabolism *Drugs*, *28*(Suppl. 1) 93, 1984

de Crombrugghe, B, *et al* · Cyclic AMP receptor protein Role in transcription activation *Science*, *224* 831, 1984

Docherty, K , and Steiner, D F Post-translational proteolysis in polypeptide hormone biosynthesis. *Annu Rev Physiol*, *44* 625, 1982

Douglass, J , *et al* Polyprotein gene expression Generation of diversity of neuroendocrine peptides *Annu Rev Biochem*, *53* 665, 1984

Feld, W S Fifty Years On Looking Back on Some Developments in Neurohumoral Physiology Mystic, CT, Lawrence Verry, 1982.

Glass, D B, and Krebs, E G Protein phosphorylation catalyzed by cyclic AMP dependent and cyclic GMP-dependent protein kinases *Annu. Rev Pharmacol Toxicol*, *20* 363, 1980

Glick, B Interrelation of the avian immune and neuroendocrine systems *J Exp Zool*, *232* 671, 1984

Godfraind, T Calcium modulators Tools for cell biology *Adv Cyclic Nucleotide Res*, *17* 551, 1984

Gustafsson, J-A, *et al* Central control of prolactin and estrogen receptors in rat liver—Expression of a novel endocrine system, the hypothalamo-pituitary-liver axis *Annu Rev Pharmacol Toxicol*, *23* 259, 1983

Hall, P F The role of the cytoskeleton in hormone action *Can J Biochem Cell Biol*, *62* 653, 1984

Imura, H , and Nakai, Y "Endorphins" in pituitary and other tissues *Annu Rev Physiol 43* 265, 1981

Kupfermann, I Role of cyclic nucleotides in excitable cells *Annu Rev Physiol*, *42* 629, 1980

McCloskey, M , and Poo, M M Protein diffusion in cell membranes Some biological implications *Int. Rev Cytol*, *87* 19, 1984

Means, A. R , and Chafouleas, J G Calmodulin in endocrine cells. *Annu Rev Physiol*, *44* 687, 1982

Means, A. R., *et al* Physiological implications of the presence, distribution, and regulation of calmodulin in eukaryotic cells. *Physiol Rev*, *62* 1, 1982

Niall, H D The evolution of peptide hormones *Annu Rev Physiol*, *44* 615 1982

Orrordan, J L Recent Advances in Endocrinology and Metabolism New York, Churchill Livingstone, 1982

Pardridge, W M Neuropeptides and the blood-brain barrier *Annu Rev Physiol*, *45* 73, 1983

Pastan, I H, and Willingham, M C Receptor-mediated endocytosis of hormones in cultured cells *Annu Rev Physiol*, *43* 239, 1981

Peck, E J, Jr Receptors for amino acids *Annu Rev Physiol*, *12* 615, 1980

Poste, G New insights into receptor regulation. *J Appl Physiol*, *57* 1297, 1984

Potemkin, V V Endocrinology Chicago, Imported Publications, 1981.

Rasmussen, H, and Barrett, P Q Calcium messenger system. An integrated view *Physiol Rev*, *64* 938, 1984

Rodbard, D (ed) Computers in Endocrinology New York, Raven Press, 1984

Roth, J , and Taylor, S I. Receptors for peptide hormones Alterations in diseases of humans *Annu Rev Physiol*, *44* 639, 1982.

Rousseau, G G Structure and regulation of the glucocorticoid hormone receptor *Mol Cell Endocrinol*, *38* 1, 1984

Sato, G H (ed) Growth of Cells in Hormonally Defined Media Cold Spring Harbor, NY, Cold Spring Harbor Institute, 1982

Sherman, M R., and Stevens, J Structure of mammalian steroid receptors Evolving concepts and methodological developments *Annu Rev Physiol*, *46* 83, 1984

Simonsen, M , *et al* Compound receptors in the cell membrane Ruminations from the borderland of immunology and physiology *Prog Allergy*, *36* 151, 1985.

Stevens, C F. Biophysical analysis of the function of receptors. *Annu. Rev Physiol*, *42* 643, 1980

Stiles, G L., *et al* β-Adrenergic receptors Biochemical mechanisms of physiological regulation *Physiol Rev*, 64 661, 1984

Tixier-Vidal, A , and Gourdji, D · Mechanism of action of synthetic hypothalamic peptides on anterior pituitary cells. *Physiol Rev*, *61* 974, 1981

Wills, M R· Laboratory Investigation of Endocrine Disorders. Woburn, MA, Butterworth, 1983

Yalow, R W Radioimmunoassay *Annu. Rev. Biophys Bioeng*, *9*.327, 1980

75

The Pituitary Hormones and Their Control by the Hypothalamus

THE PITUITARY GLAND AND ITS RELATION TO THE HYPOTHALAMUS

The *pituitary gland* (Fig. 75–1), also called the *hypophysis*, is a small gland—about 1 cm in diameter and 0.5 to 1 gram in weight—that lies in the *sella turcica* at the base of the brain and is connected with the hypothalamus by the *pituitary* (or *hypophysial*) *stalk*. Physiologically, the pituitary gland is divisible into two distinct portions the *anterior pituitary*, also known as the *adenohypophysis*, and the *posterior pituitary*, also known as the *neurohypophysis*. Between these is a small, relatively avascular zone called the *pars intermedia*, which is almost absent in the human being while much larger and much more functional in some lower animals.

Embryologically, the two portions of the pituitary originate from different sources, the anterior pituitary from *Rathke's pouch*, which is an embryonic invagination of the pharyngeal epithelium, and the posterior pituitary from an outgrowth of the hypothalamus. The origin of the anterior pituitary from the pharyngeal epithelium explains the epithelioid nature of its cells, while the origin of the posterior pituitary from neural tissue explains the presence of large numbers of glial-type cells in this gland.

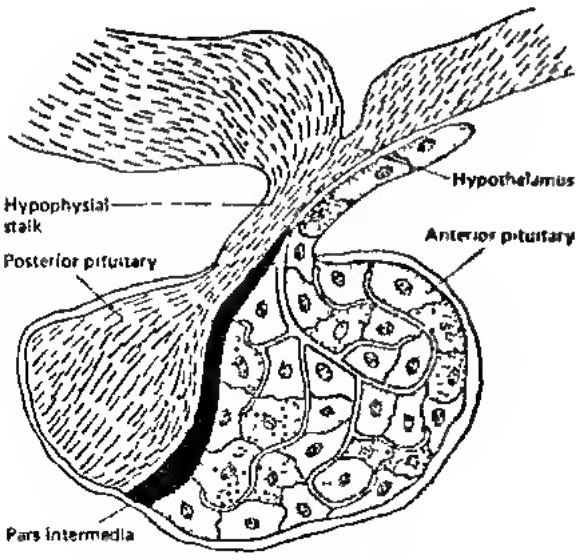


Figure 75–1. The pituitary gland

Six very important hormones plus several less important ones are secreted by the *anterior* pituitary, and two important hormones are secreted by the *posterior* pituitary. The hormones of the anterior pituitary play major roles in the control of metabolic functions throughout the body, as shown in Figure 75–2. (1) *Growth hormone* promotes growth of the animal by affecting many metabolic functions throughout the body, especially protein formation. (2) *Adrenocorticotropin* controls the secretion of some of the adrenocortical hormones, which in turn affect the metabolism of glucose, proteins, and fats. (3) *Thyroid-stimulating hormone* controls the rate of secretion of thyroxine by the thyroid gland, and thyroxine in turn controls the rates of most chemical reactions of the entire body. (4) *Prolactin* promotes mammary gland development and milk production. And two separate gonadotropic hormones, (5) *follicle-stimulating hormone* and (6) *luteinizing hormone*, control growth of the gonads as well as their reproductive activities.

The two hormones secreted by the posterior pituitary play other roles. (1) *Antidiuretic hormone* (also called *vasopressin*) controls the rate of water excretion into the urine and in this way helps control the concentration of water in the body fluids. (2) *Oxytocin* (a) helps deliver milk from the glands of the breast to the nipples during suckling, and (b) possibly helps in the delivery of the baby at the end of gestation.

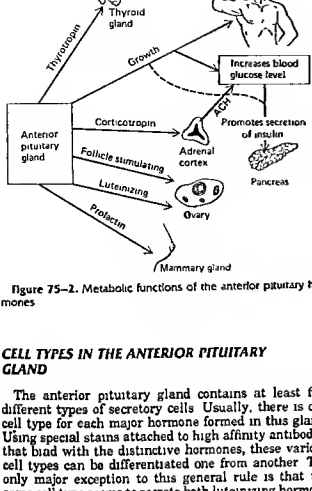


Figure 75–2. Metabolic functions of the anterior pituitary hormones

CELL TYPES IN THE ANTERIOR PITUITARY GLAND

The anterior pituitary gland contains at least five different types of secretory cells Usually, there is one cell type for each major hormone formed in this gland Using special stains attached to high affinity antibodies that bind with the distinctive hormones, these various cell types can be differentiated one from another The only major exception to this general rule is that the same cell type seems to secrete both luteinizing hormone and follicle-stimulating hormone Thus, the different cell types are

(1) *Somatotropes*—human growth hormone (hGH)

(2) *Corticotropes*—corticotropin (ACTH)

(3) *Thyrotropes*—thyroid-stimulating hormone (TSH)

(4) *Gonadotropes*—gonadotropic hormones, which include both luteinizing hormone (LH) and follicle-stimulating hormone (FSH)

(5) *Lactotropes*—prolactin (PRL)

About 30 to 40 per cent of the anterior pituitary cells are somatotropes that secrete growth hormone, and about 20 per cent are corticotropes that secrete ACTH. The other cell types each number only 3 to 5 per cent of the total; nevertheless, they secrete the extremely powerful hormones for controlling thyroid function, sexual functions, and milk secretion by the breasts.

Somatotropes stain very strongly with acid dyes and therefore are called *acidophils* Thus, pituitary tumors that secrete large quantities of human growth hormone are called *acidophilic tumors*.

The cell bodies of the cells that secrete the posterior pituitary hormones are not located in the posterior pituitary gland itself but instead are large neurons located in the hypothalamus; the hormones are then transported to the posterior pituitary glands in the axoplasm of the neurons' nerve fibers passing from the hypothalamus to the posterior pituitary gland. This will be discussed more fully later in the chapter.

CONTROL OF PITUITARY SECRETION BY THE HYPOTHALAMUS

Almost all secretion by the pituitary is controlled by either hormonal or nervous signals from the hypothalamus. Indeed, when the pituitary gland is removed from its normal position beneath the hypothalamus and transplanted to some other part of the body, its rates of secretion of the different hormones (except for prolactin) fall to low levels—in the case of some of the hormones, almost to zero.

Secretion from the posterior pituitary is controlled by nerve fibers originating in the hypothalamus and terminating in the posterior pituitary In contrast, secretion by the anterior pituitary is controlled by hormones called *hypothalamic releasing* and *inhibitory hormones* (or *factors*) secreted within the hypothalamus itself and then conducted to the anterior pituitary through minute blood vessels called *hypothalamic-hypophysial portal vessels*. In the anterior pituitary these releasing and inhibitory hormones act on the glandular cells to control their secretion. This system of control will be discussed in detail later in the chapter.

The hypothalamus receives signals from almost all possible sources in the nervous system. Thus, when a person is exposed to pain, a portion of the pain signal is transmitted into the hypothalamus. Likewise, when a person experiences some powerful depressing or exciting thought, a portion of the signal is transmitted into the hypothalamus. Olfactory stimuli denoting pleasant or unpleasant smells transmit strong signal components directly and through the amygdaloid nuclei into the hypothalamus. *Even the concentrations of nutrients, electrolytes, water, and various hormones* in the blood excite or inhibit various portions of the hypothalamus. Thus, the hypothalamus is a collecting center for information concerned with the internal well-being of the body, and in turn much of this information is used to control secretions of the many globally important pituitary hormones.

THE HYPOTHALAMIC-HYPOPHYSIAL PORTAL SYSTEM

The anterior pituitary is a highly vascular gland with extensive capillary sinuses among the glandular cells. Almost all the blood that enters these sinuses passes first through a capillary bed in the tissue of the lower tip of the hypothalamus and then through small *hypothalamic-hypophysial portal vessels* into the anterior pituitary sinuses. Thus, Figure 75–3 illustrates a small artery supplying the lowermost portion of the hypothalamus called the *median eminence* that connects inferiorly with the pituitary stalk. Small blood vessels project into the substance of the median eminence and

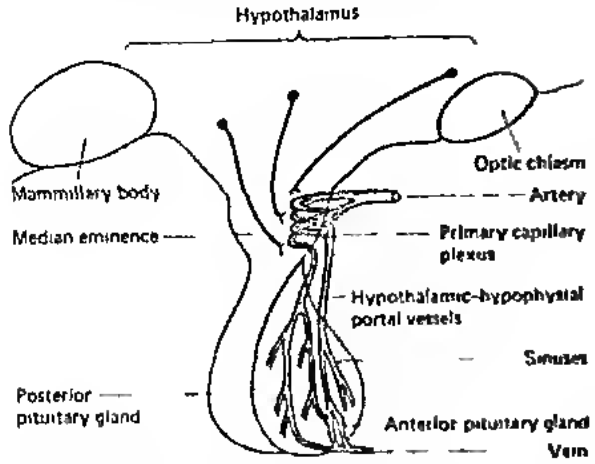


Figure 75–3. The hypothalamic hypophysial portal system

then return to its surface, coalescing to form the hypothalamic-hypophysial portal vessels These in turn pass downward along the pituitary stalk to supply blood to the anterior pituitary sinuses

Secretion of Hypothalamic Releasing and Inhibitory Hormones into the Median Eminence. Special neurons in the hypothalamus synthesize and secrete hormones called *hypothalamic releasing* and *inhibitory hormones* (or *releasing* and *inhibitory factors*) that control the secretion of the anterior pituitary hormones These neurons originate in various parts of the hypothalamus and send their nerve fibers into the median eminence and the tuber cinereum, the hypothalamic tissue that extends into the pituitary stalk. The endings of these fibers are different from most endings in the central nervous system in that their function is not to transmit signals from one neuron to another but merely to secrete the hypothalamic releasing and inhibitory hormones (factors) into the tissue fluids These hormones are immediately absorbed into the capillaries of the hypothalamic-hypophysial portal system and carried directly to the sinuses of the anterior pituitary gland.

(To avoid confusion, the student needs to know the difference between a "factor" and a "hormone" A substance that has the actions of a hormone but that has not been purified and identified as a distinct chemical compound is called a *factor*. Once it has been so identified it is thereafter known as a *hormone* instead of simply a factor.)

Function of the Releasing and Inhibitory Hormones. The function of the releasing and inhibitory hormones is to control the secretion of the anterior pituitary hormones. For each type of anterior pituitary hormone there is usually a corresponding hypothalamic releasing hormone; for some of the anterior pituitary hormones there is also a corresponding hypothalamic inhibitory factor. For most of the anterior pituitary hormones it is the releasing hormone that is important; but, for prolactin, an inhibitory hormone probably exerts most control. The hypothalamic releasing and inhibitory hormones (or factors) that are of major importance are:

1. *Thyroid-stimulating hormone releasing hormone* (TRH), which causes release of thyroid-stimulating hormone
2. *Corticotropin-releasing* (CRF), which causes release of adrenocorticotropin
3. *Growth hormone releasing hormone* (GHRH), which causes release of growth hormone, and *growth hormone inhibitory hormone* (GHIH), which is the same as the hormone *somatostatin* and which inhibits the release of growth hormone
4. *Luteinizing hormone releasing hormone* (LRH), which causes release of both luteinizing hormone and follicle-stimulating hormone—this hormone is also called *gonadotropin-releasing hormone* (GnRH)
5. *Prolactin inhibitory factor* (PIF), which causes inhibition of prolactin secretion

In addition to these more important hypothalamic hormones, still another excites the secretion of prolactin, and several hypothalamic inhibitory hormones inhibit some of the other anterior pituitary hormones Each of the more important hypothalamic hormones will be discussed in detail at the time that the specific hormonal system controlled by them is presented in this and subsequent chapters

Specific Areas In the Hypothalamus That Control Secretion of Specific Hypothalamic Releasing and Inhibitory Factors. All or most of the hypothalamic hormones are secreted at nerve endings in the median eminence before being transported to the anterior pituitary gland. Electrical stimulation of this region excites these nerve endings and therefore causes release of essentially all the hypothalamic factors. However, the neuronal cell bodies that give rise to these median eminence nerve endings are located in other discrete areas of the hypothalamus or in closely related areas of the basal brain. Unfortunately, the specific loci of the neuronal cell bodies that form the different hypothalamic releasing or inhibitory hormones are so incompletely known that it would be misleading to attempt a delineation here.

Other Hypothalamic Substances That May Have Hormonal Effects. Multiple other substances, especially many small peptides, are found in the neurons of the hypothalamus However, functions for these as hormones are only speculative. Yet, because they are of research interest they are listed here: (1) *substance P*, (2) *neurotensin*, (3) *angiotensin II*, (4) *enkephalins*, (5) *endorphins*, (6) *vasoactive inhibitory polypeptide*, and (7) *cholecystokinin-8*. Many of these same substances are also found in neurons elsewhere in the brain, suggesting that they may function as neurotransmitters both in the hypothalamus and elsewhere. In addition, some of them are in the neurons of the enteric nervous system of the gastrointestinal tract, functioning there also as neurotransmitters

and possibly as hormones released into the circulating blood from the nerve endings.

PHYSIOLOGICAL FUNCTIONS OF GROWTH HORMONE

All the major anterior pituitary hormones besides growth hormone exert their principal effects by stimulating target glands—the thyroid gland, the adrenal cortex, the ovaries, the testicles, and the mammary glands. The functions of each of these pituitary hormones are so intimately concerned with the functions of the respective target glands that, except for growth bormone, their functions and the functions of their corresponding hypothalamic releasing and inhibitory hormones will be discussed in subsequent chapters along with the target glands. Growth hormone, in contrast to other hormones, does not function through a target gland but instead exerts its effects on all or almost all tissues of the body.

Effect on Growth of Growth Hormone

Growth hormone (GH), also called *somatotropic hormone* (SH) or *somatotropin*, is a small protein molecule containing 191 amino acids in a single chein and having a molecular weight of 22,005. It causes growth of all tissues of the body that are capable of growing. It promotes both increased sizes of tbe cells end increased mitosis with development of increased numbers of cells.

Figure 75–4 illustrates weight charts of two growing rats, one of which received daily injections of growth hormone, compared with a littermate that did not receive growth hormone. This figure shows marked exacerbation of growth by growth hormone—both in the early days of life and even after the two rats had reached adulthood. In the early stages of development, all organs of tbe treated rat increased proportionately in size, but

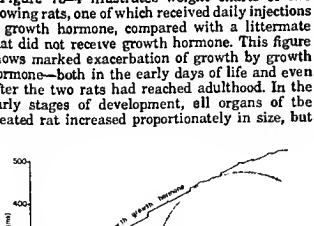


Figure 75–4. Comparison of weight gain of a rat injected daily with growth hormone with that of a normal rat.

after adulthood was reached, most of the bones ceased lengthening while the soft tissues continued to grow. This results from the fact that, once the epiphyses of the long bones have united with the shafts, further growth of bone length cannot occur even though most other tissues of the body can continue to grow throughout life.

METABOLIC EFFECTS OF GROWTH HORMONE

Aside from its general effect in causing growth, growth hormone has many specific metabolic effects as well, including especially:

(1) Increased rate of protein synthesis in all cells of the body;

(2) Increased mobilization of fatty acids from adipose tissue, and increased use of the fatty acids for energy; and

(3) Decreased rate of glucose utilization throughout the body.

Thus, in effect, growth hormone enhances the body protein, uses up the fat stores, and conserves carbohydrate. It is probable that the increased rate of growth results mainly from the increased rate of protein synthesis.

Role of Growth Hormone in Promoting Protein Deposition

Although the most important cause of the increased protein deposition caused by growth hormone is not known, a series of different effects are known, all of which can lead to enhanced protein. These effects are:

1. Enhancement of Amino Acid Transport Through the Cell Membranes. Growth hormone directly enhances transport of at least some and perhaps most amino acids through the cell membranes to the interior of the cells. This increases the concentrations of the amino acids in the cells and is presumed to be at least partly responsible for the increased protein synthesis. This control of amino acid transport is similar to the effect of insulin in controlling glucose transport through the membrane, as discussed in Chapters 67 and 78.

2. Enhancement of Protein Synthesis by the Ribosomes. Even when the amino acids are not increased in the cells, growth hormone still causes protein to be synthesized in increased amounts in the cells. This is believed to be caused partly by a direct effect on the ribosomes, making them produce greater numbers of protein molecules; the mechanism by which this effect occurs is as yet unknown.

3. Increased Transcription of DNA to Form RNA. Over more prolonged periods of time (24 to 48 hours), growth hormone also stimulates the transcription of DNA in the nucleus, causing formation of increased quantities of RNA. This in turn promotes protein synthesis and also promotes

growth if sufficient energy, amino acids, vitamins, and other necessities for growth are available. In the long run, this perhaps is the most important of all the functions of growth hormone.

A. Decreased Catabolism of Protein and Amino Acid. In addition to the increase in protein synthesis, there is a decrease in the breakdown of cell protein. A possible if not probable reason for this is that growth hormone also mobilizes large quantities of free fatty acids from the adipose tissue, and these in turn are used to supply most of the energy for the body cells, thus acting as a potent "protein sparer."

Summary. Growth hormone enhances almost all facets of amino acid uptake and protein synthesis by cells, while at the same time reducing the breakdown of proteins.

Effect of Growth Hormone in Enhancing Fat Utilization for Energy

Growth hormone has a specific effect, but one that takes several hours to develop, in causing release of fatty acids from adipose tissue and, therefore, increasing the fatty acid concentration in the body fluids. In addition, in the tissues it enhances the conversion of fatty acids to acetyl-CoA with subsequent utilization of this for energy. Therefore, under the influence of growth hormone, fat is utilized for energy in preference to both carbohydrates and proteins.

Some research workers have considered this fat mobilization effect of growth hormone to be one of its most important functions and have also considered the protein-sparing effect to be a major factor that promotes protein deposition and growth. However, growth hormone mobilization of fat requires hours to occur, whereas enhancement of cellular protein synthesis can begin in minutes under the influence of growth hormone

Ketogenic Effect of Growth Hormone. Occasionally, fat mobilization under the influence of excessive amounts of growth hormone is so great that excessive quantities of acetoacetic acid are formed by the liver and are released into the body fluids, thus causing *ketosis*. This excessive mobilization of fat from the adipose tissue also frequently causes a fatty liver.

Effect of Growth Hormone on Carbohydrate Metabolism

Growth hormone has three major effects on cellular metabolism of glucose: (1) decreased utilization of glucose for energy, (2) enhancement of glycogen deposition in the cells, and (3) diminished uptake of glucose by the cells.

Decreased Use of Glucose for Energy. Unfortunately, we do not know the precise mechanism by which growth hormone decreases glucose utili- by the cells However, the decrease probably 'ts partially from the increased mobilization and utilization of fatty acids for energy caused by growth hormone. That is, the fatty acids form large quantities of acetyl-CoA that in turn initiate feedback effects to block the glycolytic breakdown of glucose and glycogen.

Enhancement of Glycogen Deposition. Since glucose and glycogen cannot be utilized for energy, the glucose that does enter the cells is rapidly polymerized into glycogen and deposited. Therefore, the cells rapidly become saturated with glycogen and can store no more.

Diminished Uptake of Glucose by the Cells and Increased Blood Glucose Concentration. When growth hormone is first administered to an animal, the cellular uptake of glucose is enhanced, and the blood glucose concentration falls slightly. However, this effect lasts for only 30 minutes to an hour or so and is then superseded by exactly the opposite effect, namely decreased transport of glucose through the cell membrane. This perhaps results from the fact that the cells have already taken up an excess of glucose that they are having difficulty using Without normal cellular uptake and use, the blood concentration of glucose increases, sometimes to as high as 50 to 100 per cent above normal.

Necessity of Insulin and Carbohydrate for the Growth-Promoting Action of Growth Hormone. Growth hormone fails to cause growth in an animal lacking a pancreas, and it also fails to cause growth if carbohydrates are excluded from the diet. This shows that adequate insulin activity as well as adequate availability of carbohydrates is necessary for growth hormone to be effective. Part of this requirement for carbohydrates and insulin is to provide the energy needed for the metabolism of growth. But there seem to be other effects as well. Especially important is one effect of insulin on enhancing amino acid transport into cells in the same way that it enhances glucose transport

Diabetogenic Effect of Growth Hormone. We have already pointed out that growth hormone leads to moderately increased blood glucose concentration. This in turn stimulates the beta cells of the islets of Langerhans to secrete extra insulin In addition to this effect, growth hormone also has a slight direct stimulatory effect on the beta cells as well. The combination of these two effects sometimes so greatly overstimulates insulin secretion by the beta cells that they literally "burn out." When this occurs the person develops diabetes mellitus, a disease that will be discussed in detail in Chapter 78 Therefore, growth hormone is said to have a *diabetogenic effect*

Diabetogenic Effects of Other Anterior Pituitary Hormones. Growth hormone is not the only anterior pituitary hormone that increases the blood glucose concentration. At least three others can do the same adrenocorticotropin, thyroid stimulating hormone, and prolactin. Especially important is adrenocorticotropin, which increases the rate of cortisol secretion by the adrenal cortex. Cortisol then increases the blood glucose concentration by increasing the rate of gluconeogenesis This effect, quantitatively, is probably equally as diabetogenic as the effect of growth hormone.

Pituitary Diabetes. From the preceding discussion it is readily apparent that specific increase in secretion of growth hormone or generalized increase in secretion of all the anterior pituitary hormones causes elevated *blood glucose concentration; this condition is called pituitary diabetes,* and it differs from *diabetes mellitus,* which results from insulin lack, in the following ways: First, in pituitary diabetes the rate of glucose utilization by the cells is only moderately depressed, in comparison with almost no utilization in diabetes mellitus. Second, the blood glucose concentration is relatively *refractory to insulin*—that is, it decreases very little in response to an insulin injection because adequate insulin is already available in the body. Third, many of the side effects that result from reduced carbohydrate metabolism in diabetes mellitus are absent in pituitary diabetes.

STIMULATION OF CARTILAGE AND BONE GROWTH—ROLE OF THE SOMATOMEDINS

Though we have discussed at length the role of growth hormone in causing growth of an animal, growth hormone does not have a *direct* effect on the growth of the skeletal elements cartilage and bone. For instance, when growth hormone is applied directly to cartilage chondrocytes cultured outside the body, no discernible proliferation or enlargement of the chondrocytes occurs, even though growth hormone injected into the intact animal does cause such proliferation and growth. To make a long story short, it has been found that growth hormone acts indirectly on cartilage and bone by causing the liver to form several small proteins called *somatomedins,* with molecular weights varying between 4500 and 7500. These somatomedins then act on the cartilage and bone to promote their growth. One or more of the somatomedins are required for deposition of chondroitin sulfate and collagen, both of which are necessary for cartilage and bone growth.

At least four different somatomedins have been isolated. However, all seem to have closely similar actions, differing in potency.

Once the epiphyses of the long bones have united with the shafts, the bones can no longer increase in length but they can continue to increase in thickness because of growth beneath the periosteum. Therefore, excessive growth hormone after adolescence cannot cause further increase in height of a person. However, it can cause disproportionate growth of the thickness of the long bones as well as continued growth of the membranous bones, which have no epiphyses to unite. For instance, the jawbones can still be stimulated to grow, causing forward protrusion of the chin and lower teeth. Likewise the bones of the skull wall continue to grow in thickness, giving rise to bony protrusions over the eyes.

Other Metabolic and Growth Functions of Somatomedin. Though growth hormone can cause all the metabolic effects that we have discussed thus far, except cartilage and bone growth, when added directly to tissues removed from the body, the concentrations of growth hormone that are required to cause these actions are often many times as great as those known to exist in the body. Furthermore, experiments have shown that the different somatomedins in very small concentrations can cause essentially all the same effects as those caused by the direct actions of growth hormone. Therefore, it is probable that most of the metabolic functions of growth hormone are caused not by its direct effects on the tissues but instead indirectly through the somatomedins.

REGULATION OF GROWTH HORMONE SECRETION

For many years it was believed that growth hormone was secreted primarily during the period of growth but then disappeared from the blood at adolescence. However, this has proved to be very far from the truth, because after adolescence, secretion continues at a rate almost as great as that in childhood. Furthermore, the rate of growth hormone secretion increases and decreases within minutes, as illustrated in Figure 75–5, sometimes for reasons that are not et all understood, but at other times definitely in relation to the person's state of nutrition or stress, such as during (1) *starvation,* (2) *hypoglycemia* or *low concentration of fatty acids in the blood,* (3) *exercise,* (4) *excitement,* and (5) *trauma.* And it characteristically increases during the first two hours of *deep sleep,* as also is shown in Figure 75–5.

The normal concentration of growth hormone in the plasma of an adult is about 3 millimicrograms per milliliter and in a child about 5 millimicrograms per milliliter. However, these values often increase to as high as 50 millimicrograms per milliliter after depletion of the body stores of proteins or carbohydrates. Under acute conditions, hypoglycemia is a far more potent stimulator of growth hormone secretion than is an acute decrease in protein intake. In fact, acute administration of some amino acids, especially *arginine,* can cause acute *increase* in growth hormone secretion;

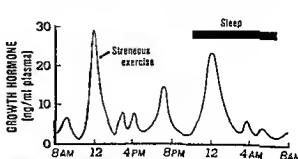


Figure 75–5. Typical variations in growth hormone secretion throughout the day, illustrating the especially powerful effect on secretion caused by strenuous exercise, and illustrating also the high rate of growth hormone secretion that occurs during the first few hours of deep sleep

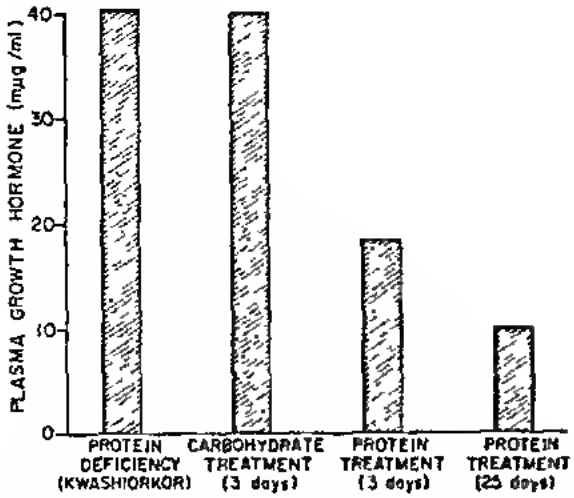


Figure 75–6. Effect of extreme protein deficiency on the concentration of growth hormone in the plasma in the disease kwashiorkor. The figure also shows the failure of carbohydrate treatment but the effectiveness of protein treatment in lowering growth hormone concentration. (Drawn from data in Pimstone *Am J Clin Nutr*, *21* 482, 1968.)

this hormone then increases the rate of protein synthesis.

On the other hand, in chronic conditions the degree of cellular protein depletion seems to correlate more with the level of growth hormone secretion than with the availability of glucose. For instance, the extremely high levels of growth hormone that occur during starvation are very closely related to the amount of protein depletion. Figure 75–6 illustrates this relationship: the first column shows growth hormone levels in children with extreme protein deficiency during the malnutrition disease called *kwashiorkor*; the second column shows the levels in the same children after 3 days of treatment with more than adequate quantities of carbohydrates in their diets, illustrating that the carbohydrates did not lower the plasma growth hormone concentration; the third and fourth columns show the levels after treatment with protein supplement in their diet for 3 and 25 days, with concomitant decrease in the hormone. These results demonstrate that under very severe conditions of protein malnutrition, adequate calories alone are not sufficient to correct the excess production of growth hormone. Instead, the protein deficiency must also be corrected before the growth hormone concentration will return to normal.

Role of the Hypothalamus, Growth Hormone Releasing Hormone, and Somatostatin in the Control of Growth Hormone Secretion

From the preceding description of the many different factors that can affect growth hormone secretion, one can readily understand the perplexity that has faced physiologists in their attempt to unravel the mysteries of the regulation of growth hormone secretion. However, it is known that growth hormone secretion is controlled almost entirely in response to two factors secreted in the hypothalamus and then transported to the anterior pituitary gland through the hypothalamic-hypophysial portal vessels. These are *growth hormone releasing hormone* (GHRH) and *growth hormone inhibitory hormone* (GHIH), also called *somatostatin*.

The hypothalamic nucleus that causes secretion of growth hormone releasing hormone is the ventromedial nucleus, the same area of the hypothalamus that is known to be sensitive to hypoglycemia and to cause hunger in hypoglycemic states. The secretion of somatostatin is controlled by other nearby areas of the hypothalamus. Therefore, it is reasonable to believe that some of the same signals that modify a person's behavioral feeding instincts also alter the rate of growth hormone secretion. In a similar manner, hypothalamic signals depicting emotions, stress, and trauma can all affect hypothalamic control of growth hormone secretion. In fact, definitive experiments have shown that catecholamines, dopamine, and serotonin, each of which is released by a different neuronal system in the hypothalamus, will increase the rate of growth hormone secretion.

Most of the control of growth hormone secretion is probably mediated through growth hormone releasing hormone rather than through the inhibitory hormone somatostatin. However, it is important to note that somatostatin is also secreted by the delta cells of the islets of Langerhans in the pancreas, and that it can inhibit the secretion of insulin and glucagon by the beta and alpha cells in the islets of Langerhans in the same way that it inhibits the secretion of growth hormone by the anterior pituitary gland. For this reason, somatostatin might have a widespread role in modulating the functions of multiple hormonal systems.

When growth hormone is administered to an animal over a period of hours, the rate of endogenous growth hormone secretion becomes decreased. This illustrates that growth hormone secretion, as is true of essentially all other hormones, is subject to typical negative feedback control. However, the nature of this feedback mechanism and whether it is mediated through inhibition of growth hormone releasing hormone or enhancement of somatostatin are still unknown.

In summary, our present knowledge of the regulation of growth hormone secretion is not sufficient to describe a composite picture. Yet, because of its undoubted long-term effect on promoting protein synthesis and growth of tissues, one would like to propose that the major long-term controller of growth hormone secretion is the state of nutrition of the tissues themselves, especially their level of protein nutrition. That is, nutritional deficiency or excess tissue need for cellular proteins—for instance, after a severe bout of exercise when the

muscles' nutritional status has been taxed—would in some way increase the rate of growth hormone secretion. The growth hormone in turn would promote synthesis of new proteins while at the same time conserving the protein already present in the cells.

ABNORMALITIES OF GROWTH HORMONE SECRETION

Panhypopituitarism. This term means decreased secretion of all the anterior pituitary hormones. The decrease in secretion may be congenital (present from birth), or it may occur suddenly or slowly at any time during the life of the individual.

Dwarfism. Some instances of dwarfism result from deficiency of anterior pituitary secretion during childhood. In general, the features of the body develop in appropriate proportion to each other, but the rate of development is greatly decreased. A child who has reached the age of 10 years may have the bodily development of a child of 4 to 5 years, whereas the same person on reaching the age of 20 years may have the bodily development of a child of 7 to 10 years.

The dwarf usually does not exhibit specific thyroid deficiency or adrenocortical deficiency, for the entire body remains so small that only small quantities of thyroid-stimulating and adrenocorticotropic hormones are needed. Also, there is no mental retardation. On the other hand, the panhypopituitary dwarf does not pass through puberty and never secretes a sufficient quantity of gonadotropic hormones to develop adult sexual functions. In one third of the dwarfs, however, the deficiency is of growth hormone alone; these individuals do mature sexually and occasionally do reproduce. In a very rare condition (the Lorain dwarf), the rate of growth hormone secretion is actually normal or high but there is a hereditary inability to form somatomedin in response to the growth hormone.

Treatment with Human Growth Hormone. The growth hormones of different species of animals are sufficiently different from each other that they will function to cause growth only in the one animal species or at most very closely allied species. For this reason, growth hormone prepared from lower animals (except to some extent from primates) is not effective in human beings. Therefore, the growth hormone of the human being is called *human growth hormone* (hGH) to distinguish it from the others.

Unfortunately, in the past it has been very difficult to obtain sufficient quantities of human growth hormone to treat patients with growth hormone deficiency except on an experimental basis because it has had to be prepared from human pituitary glands. However, human growth hormone can now be synthesized by *Escherichia coli* bacteria as a result of successful application of recombinant DNA technology. Therefore, this hormone is now beginning to become available in sufficient quantities for treatment purposes. Dwarfs who have pure growth hormone deficiency can be completely cured. Also, human growth hormone might prove to be beneficial in other metabolic disorders, because of its widespread metabolic functions.

Panhypopituitarism in the Adult. Panhypopituitarism occurring in adulthood frequently results from one of three common abnormalities. Two tumorous conditions, craniopharyngiomas and chromophobe tumors, may compress the pituitary gland until the functioning anterior pituitary cells are totally or almost totally destroyed. The third cause is thrombosis of the pituitary blood vessels. This occurs occasionally when a mother develops circulatory shock following birth of a baby.

The effects of adult panhypopituitarism, in general, are (1) hypothyroidism, (2) depressed production of glucocorticoids by the adrenal glands, and (3) suppressed secretion of the gonadotropic hormones to the point that sexual functions are lost. Thus, the picture is that of a lethargic person (from lack of thyroid hormones) who is gaining weight because of lack of fat mobilization by growth, adrenocorticotropic, adrenocortical, and thyroid hormones and who has lost all sexual functions. Except for the abnormal sexual functions, the patient can usually be treated satisfactorily by administration of adrenocortical and thyroid hormones.

Gigantism. Occasionally, the acidophilic, growth hormone–producing cells of the anterior pituitary become excessively active, and sometimes even acidophilic tumors occur in the gland. As a result, large quantities of growth hormone are produced. All body tissues grow rapidly, including the bones, and if the condition occurs before adolescence, that is, before the epiphyses of the long bones have not become fused with the shafts, height increases so that the person becomes a giant with heights as great as 8 to 9 feet.

The giant ordinarily has *hyperglycemia*, and the beta cells of the islets of Langerhans in the pancreas are prone to degenerate, partially because they become overactive owing to the hyperglycemia and partially because of a direct overstimulating effect of growth hormone on the islet cells. Consequently, about 10 per cent of the giants finally *develop full-blown diabetes mellitus*.

Most giants, unfortunately, eventually develop panhypopituitarism if they remain untreated, because the gigantism is usually caused by a tumor of the pituitary gland that grows until the gland itself is destroyed. This general deficiency of pituitary hormones usually causes death in early adulthood. However, once gigantism is diagnosed, further development can often be blocked by microsurgical removal of the tumor from the pituitary gland or, if not in this way, by irradiation of the gland.

Acromegaly. If an acidophilic tumor occurs after adolescence—that is, after the epiphyses of the long bones have fused with the shafts—the person cannot grow taller; but the soft tissues can continue to grow, and the bones can grow in thickness. This condition, which is illustrated in Figure 75–7, is known as *acromegaly*. Enlargement is especially marked in the small bones of the hands and feet and in the *membrinous bones*, including the cranium, the nose, the bosses on the forehead, the supraorbital ridges, the lower jawbone, and portions of the vertebrae, for their growth does not cease at adolescence. Consequently, the jaw protrudes forward, sometimes as much as a half inch, the forehead slants forward because of excess development of the supraorbital ridges, the nose increases to as much as twice normal size, the foot requires a size 14 or larger shoe, and the fingers become extremely thickened so that the hand develops a size almost twice normal. In addition to these effects, changes in the vertebrae ordi-

Figure 75–7. An acromegalic patient. (Courtesy of Dr. Herbert Langford.)

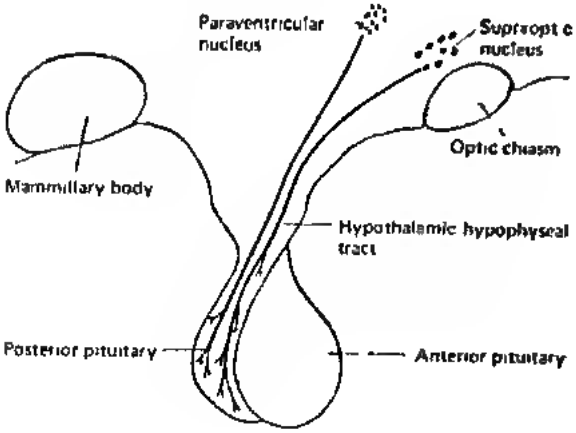


Figure 75–8. Hypothalamic control of the posterior pituitary.

narily cause a hunched back, which is known clinically as *kyphosis*. Finally, many soft tissue organs, such as the tongue, liver, and especially the kidneys, become greatly enlarged.

THE POSTERIOR PITUITARY GLAND AND ITS RELATION TO THE HYPOTHALAMUS

The *posterior pituitary gland*, also called the *neurohypophysis*, is composed mainly of glial-like cells called *pituicytes*. However, the pituicytes do not secrete hormones; they act simply as a supporting structure for large numbers of *terminal nerve fibers* and *terminal nerve endings* from nerve tracts that originate in the *supraoptic* and *paraventricular nuclei* of the hypothalamus, as shown in Figure 75–8. These tracts pass to the neurohypophysis through the *pituitary stalk* (hypophysial stalk). The nerve endings are bulbous knobs that lie on the surfaces of capillaries onto which they secrete the two posterior pituitary hormones: (1) *antidiuretic hormone* (ADH), also called *vasopressin*, and (2) *oxytocin*.

If the pituitary stalk is cut near the pituitary gland, leaving the entire hypothalamus intact, the posterior pituitary hormones continue, after a transient decrease for a few days, to be secreted almost normally, but they are then secreted by the cut ends of the fibers within the hypothalamus and not by the nerve endings in the posterior pituitary. The reason for this is that the hormones are initially synthesized in the cell bodies of the supraoptic and paraventricular nuclei and are then transported in combination with "carrier" proteins called *neurophysins* down to the nerve endings in the posterior pituitary gland, requiring about a week to reach the gland.

ADH is formed primarily in the supraoptic nuclei, whereas oxytocin is formed primarily in the paraventricular nuclei. However, each of these two nuclei can synthesize approximately one sixth as much of the second hormone as of its primary hormone.

Under resting conditions, large quantities of both ADH and oxytocin accumulate in large secretory granules in the nerve endings of the posterior pituitary gland, still loosely bound with their respective neurophysins. Then, when nerve impulses are transmitted downward along the fibers from the supraoptic or paraventricular nuclei, the hormone is immediately released from the nerve endings by the usual secretory mechanism of *exocytosis* and is absorbed into adjacent capillaries. Both the neurophysin and the hormone are secreted together, but since they are only loosely bound to each other, it is believed that the hormone separates almost immediately. The neurophysin has no known function after leaving the nerve terminals.

CHEMICAL NATURE OF ANTIDIURETIC HORMONE (VASOPRESSIN) AND OXYTOCIN

Both oxytocin and ADH (vasopressin) are polypeptides containing nine amino acids. The amino acid sequences of these are the following:

Vasopressin:
Cys-Tyr-Phe-Gln-Asn-Cys-Pro-Arg-GlyNH$_2$

Oxytocin:
Cys-Tyr-Ile-Gln-Asn-Cys-Pro-Leu-GlyNH$_2$

Note that these two hormones are almost identical except that in vasopressin phenylalanine and arginine replace isoleucine and leucine of the oxytocin molecule. The similarity of the molecules explains their occasional functional similarities.

PHYSIOLOGICAL FUNCTIONS OF ANTIDIURETIC HORMONE (VASOPRESSIN)

Extremely minute quantities of antidiuretic hormone (ADH)—as small as 2 millimicrograms—when injected into a person can cause *antidiuresis*, that is, decreased excretion of water by the kidneys. This antidiuretic effect was discussed in detail in Chapter 36. Briefly, in the absence of ADH, the collecting ducts and tubules are almost totally impermeable to water, which prevents significant reabsorption of water and therefore allows extreme loss of water into the urine. On the other hand, in the presence of ADH the permeability of the collecting ducts and tubules to water increases greatly and allows most of the water to be reabsorbed as the tubular fluid passes through these ducts, thereby conserving water in the body.

The precise mechanism by which ADH acts on the ducts to increase their permeability is unknown. However, the hormone first becomes fixed to specific receptors on the basal side of the collecting duct epithelial cells, causing them to form large quantities of cyclic AMP in their cytoplasm. This in turn acts on the membranes of the luminal side of the cells, opening many pores that allow free diffusion of water between the tubular and peritubular fluids, though the manner in which cyclic AMP acid causes its effect on the pores is mainly unknown. The water is then absorbed from the collecting duct (and tubule) by osmosis, as was explained in relation to the concentrating mechanism of the kidney in Chapter 36.

Regulation of ADH Production

Osmotic Regulation. When a concentrated electrolyte solution is injected into the artery supplying the hypothalamus, the ADH neurons in the supraoptic and paraventricular nuclei immediately transmit impulses into the posterior pituitary to release large quantities of ADH into the circulating blood. Conversely, injection of pure

water into this artery causes complete cessation of the impulses and essentially complete cessation of ADH secretion. The ADH that is already in the tissues is destroyed at a rate of approximately one half every 15 to 20 minutes. Thus, the concentration of ADH in the body fluids can change from small amounts to large amounts, or vice versa, in only a few minutes.

It has been postulated that neurons that are separate from the ADH neurons but located near them function as *osmoreceptors*. These presumably increase and decrease in size in relation to the degree of concentration of the extracellular fluid. When the extracellular fluid is highly dilute, osmosis of water into the cell supposedly increases its volume, while concentrated extracellular fluid supposedly reduces the volume. Regardless of the mechanism, concentrated body fluids do stimulate *the supraoptic nuclei, whereas dilute body fluids* inhibit them. Therefore, a feedback control system is available to control the total osmotic pressure of the body fluids, operating as follows:

When the body fluids become highly concentrated, the supraoptic nuclei become excited, impulses are transmitted to the posterior pituitary, and ADH is secreted. This passes by way of the blood to the kidneys, where it increases the permeability of the collecting ducts to water. As a result, most of the water is then reabsorbed from the tubular fluid, while electrolytes continue to be lost into the urine. This dilutes the extracellular fluid, returning it to a reasonably normal osmotic composition.

Role of ADH in Controlling Extracellular Fluid Sodium Ion Concentration. Under normal conditions, about 95 per cent of the total osmotic pressure of the extracellular fluids is determined by the sodium ion concentration of these fluids (because of the high concentration of sodium and because the anion concentration "follows" the sodium concentration). Therefore, in effect, when one says that ADH controls the osmolality of the extracellular fluids, this also means that ADH controls the sodium ion concentration of these fluids at the same time. Furthermore, recent experiments have shown that the major change that occurs in the composition of the extracellular fluid following either increased or decreased ADH secretion is a change in sodium ion concentration, with much less change in volume of water in the extracellular fluids. Therefore, it is becoming clear that ADH is a very potent controller of sodium ion concentration, a fact that was discussed in detail in Chapter 36.

Stimulation of ADH Secretion by Low Blood Volume—Pressor Effect of ADH. ADH in moderate to high concentrations has a very potent effect of constricting the arterioles and therefore of increasing the arterial pressure. Also, one of the most powerful stimuli of all for increasing the secretion of ADH is severe loss of blood volume. As little as 10 per cent loss of blood will promote a moderate increase in ADH secretion, and 25 per cent or more blood loss can cause as much as 50 times normal rates of secretion.

The increased secretion is believed to result mainly from the low pressure caused in the atria of the heart by the low blood volume. The relaxation of the atrial stretch receptors supposedly elicits the increase in ADH secretion. However, the baroreceptors of the carotid, aortic, and pulmonary regions also participate in the control of ADH secretion.

It is generally stated that the normal concentration of ADH in the body fluids, between 1 and 2 picograms per milliliter (trillionths of a gram per milliliter), is too small to cause any significant pressor effect. However, recent studies by Cowley in which the compensatory role of the nervous reflexes was eliminated showed that between 5 and 10 mm Hg of the normal arterial pressure is maintained by this normal ADH. Also, he showed that, when the arterial pressure falls, the amount of ADH secreted is enough to bring the pressure three fourths of the way back to normal. This shows that ADH plays a much greater role in arterial pressure homeostasis than has previously been believed.

Because ADH has this potent pressor effect, it is also called vasopressin.

Other Factors That Affect ADH Production. Other factors that frequently increase the output of ADH include *trauma* to the body, *pain*, *anxiety*, and drugs such as *morphine*, *nicotine*, *tranquilizers*, and some *anesthetics*. Each of these factors can cause retention of water in the body. This probably explains the frequent accumulation of water in many emotional states, and possibly also explains the diuresis that occurs when the state is over.

A substance that inhibits ADH secretion is *alcohol*. Therefore, during an alcoholic bout, lack of ADH allows marked diuresis. Alcohol probably also dilates the afferent arterioles of the nephrons, which adds to its diuretic effect.

Diabetes Insipidus. Diabetes insipidus is the disease that occurs when the supraopticohypophysial system secreting ADH fails. This will occur only when the ADH neurons in the supraoptic and paraventricular nuclei are destroyed. It will not occur when the posterior pituitary gland alone is destroyed or when the pituitary stalk is cut, because the cut nerve fibers can continue to secrete ADH. In a person with full-blown diabetes insipidus, lack of ADH makes the urine almost always very dilute except in very severe dehydration states. The urine specific gravity remains almost constantly between 1.002 and 1.006, the urine output is usually 4 to 6 liters per day but can be as great as 12 to 15 liters per day, depending principally on how much water the person drinks. Furthermore, the rapid loss of fluid in the urine creates a constant thirst, which keeps the water flushing through the body.

The person with diabetes insipidus has a tendency to become dehydrated. However, this tendency is usually quite well offset by the increased thirst. Under conditions of circulatory stress or when water might not be adequately available, the fluid loss can become serious.

Diabetes insipidus occurs most frequently as a result of a tumor of the hypothalamus or hypophysis that destroys the portions of the hypothalamus that control ADH secretion. It can be treated easily by simply injecting vasopressin suspended in oil (for slow release) once every two days.

Excess Secretion of ADH—Syndrome of Inappropriate ADH Secretion. Occasionally, excess ADH is secreted by the hypophysial-posterior pituitary system

or, much more usually by certain types of tumors in the body, particularly bronchogenic carcinomas of the lungs. When this occurs, a condition called *syndrome of inappropriate ADH secretion* develops, characterized by greatly decreased sodium ion concentration in the extracellular fluid but by only a few per cent increase in body water. The reasons for these effects were discussed in detail in Chapter 36; they especially point up the fact that ADH is much more a controller of sodium ion concentration than of body fluid volume.

Other Actions of ADH

Large doses of ADH can also cause contraction of almost any smooth muscle tissue in the body, including contractions of most of the intestinal musculature, the bile ducts, and the uterus However, the concentrations required to cause these effects are far greater than those required to cause antidiuresis, and it is doubtful that these are significant physiological effects for normal function of the body.

Some psychologists have suggested that ADH secreted into the hypothalamus itself or into nearby areas of the brain increases one's memory retention

OXYTOCIC HORMONE

Effect on the Uterus. An oxytocic substance is one that causes contraction of the pregnant uterus. The hormone *oxytocin*, in accordance with its name, powerfully stimulates the pregnant uterus, especially so toward the end of gestation. Therefore, many obstetricians believe that this hormone is at least partially responsible for effecting birth of the baby. This is supported by the following facts: (1) In a hypophysectomized animal, the duration of labor is prolonged, thus indicating a possible effect of oxytocin during delivery. (2) The amount of oxytocin in the plasma increases during labor, especially during the last stage (3) Stimulation of the cervix in a pregnant animal elicits nervous signals that pass to the hypothalamus and cause increased secretion of oxytocin. These effects and this possible mechanism for aiding the birth of the baby will be discussed in Chapter 82.

Effect of Oxytocin on Milk Ejection. Oxytocin plays an especially important role in the process of lactation, a role that is far more *certain* than its possible role in delivery of the baby. In lactation, it causes milk to be expressed from the alveoli into the ducts so that the baby can obtain it by suckling. This mechanism works as follows: The suckling stimuli on the nipple of the breast cause signals to be transmitted through the sensory nerves to the brain. The signals pass upward through the brain stem and finally reach the oxytocin neurons in the paraventricular and supraoptic nuclei in the hypothalamus, to cause release of oxytocin. The oxytocin then is carried by the blood to the breasts where it causes contraction of *myoepithelial cells* that lie outside of and form a latticework surrounding the alveoli of the mammary glands. In less than a minute after the beginning of suckling, milk begins to flow. Therefore, this mechanism is frequently called *milk letdown* or *milk ejection*. This process is discussed further in Chapter 82 in relation to lactation.

REFERENCES

Almers, W, and Stirling, C Distribution of transport proteins over animal cell membranes. *J Membr. Biol*, 77.169, 1984

Arad, Z, and Skadhauge, E: Plasma hormones (arginine vasotocin, prolactin, aldosterone, and corticosterone) in relation to hydration state, NaCl intake, and egg laying in fowls. *J Exp Zool*, 232.707, 1984

Assenmacher, I, and Farmer, D S (eds) Environmental Endocrinology New York, Springer-Verlag, 1978

Austin, C R, and Short, R V (eds) Mechanics of Hormone Action New York, Cambridge University Press, 1979

Barrington, E J W (ed.) Hormones and Evolution. New York, Academic Press, 1979

Baxter, J D, and MacLeod, K M Molecular basis for hormone action In Bondy, P K, and Rosenberg L. E. (eds) Metabolic Control and Disease, 8th Ed. Philadelphia, W B Saunders Co , 1980, p. 104

Bizollon, C A (ed) Radioimmunology, 1979 New York, Elsevier North-Holland, 1979

Bowers, C Y, *et al* Hypothalamic peptide hormones. Chemistry and physiology *In* DeGroot, L J, *et al* (eds) Endocrinology Vol 1 New York, Grune & Stratton, 1978, p 65

Brownstein, M J Biosynthesis of vasopressin and oxytocin *Annu Rev Physiol*, *45* 129, 1983

Chawla, R K, *et al* Structural variants for human growth hormone Biochemical, genetic, and clinical aspects *Annu Rev Med* *34* 519, 1983

Cheung, W Y Calmodulin its potential role in cell proliferation and heavy metal toxicity *Fed Proc*, *43*.2995, 1984

Chrousos, G P, *et al* NIH Conference Clinical applications of corticotropin-releasing factor *Ann Intern Med*, *102* 344, 1985

Conn, P M, *et al* Gonadotropin-releasing hormone Molecular and cell biology, physiology, and clinical applications *Fed Proc*, *43* 2351, 1984

Daly, J W Forskolin, adenylate cyclase, and cell physiology An overview *Adv Cyclic Nucleotide Res*, *17* 81, 1984

DeGroot, L J (ed) Endocrinology, New York Grune & Stratton, 1979.

Dillon, R S Handbook of Endocrinology Diagnosis and Management of Endocrine and Metabolic Disorders, 2nd Ed Philadelphia, Lea & Febiger, 1980

Doris, P A Vasopressin and central integrative processes. *Neuroendocrinology*, *38* 75, 1984

Eazm, C, *et al* (eds) Pituitary Diseases Boca Raton, Fla , CRC Press, 1980

Fiddes, J C, and Talmadge, K Structure, expression, and evolution of the genes for the human glycoprotein hormones *Recent Prog Horm Res*, *40* 43, 1984

Gann, D S, *et al* Neural interaction in control of adrenocorticotropin *Fed Proc*, *44* 161, 1985

Gray, C H, and James, V H T Hormones in Blood, New York, Academic Press, 1979

Handler, J S, and Orloff, J Antidiuretic hormone *Annu Rev Physiol*, *43* 611, 1981

Hershman, J M (eds) Management of Endocrine Disorders. Philadelphia, Lea & Febiger, 1980

Horto-Delarche, F Glucocorticoid receptors and steroid sensitivity in normal and neoplastic human lymphoid tissues. A review *Cancer Res*, *44* 431, 1984

Hsueh, A J W, and Jones, P B C Gonadotropin releasing hormone Extrapituitary actions and paracrine control mechanisms. *Annu Rev Physiol*, *45* 83, 1983.

Jeffcoate, S. L., and Hutchinson, J S. M (eds) The Endocrine Hypothalamus. New York, Academic Press, 1978

Johnston, F E., *et al* (eds) Human Physical Growth and Maturation Methodologies and Factors. New York, Plenum Press, 1980.

Keller-Wood, M E., and Dallman, M. F. Corticosteroid inhibition of ACTH secretion. *Endocr. Rev*, *5* 1, 1984

Kleeman, C R, and Bert, T: The neurohypophysial hormones Vasopressin *In* DeGroot, L. J., *et al* (eds) Endocrinology Vol 1, New York, Grune & Stratton, 1979, p 253

Kostyo, J L., and Isaksson, O Growth hormone and the regulation of somatic growth. *Int Rev Physiol*, *13* 255, 1977

Krulich L.: Central neurotransmitters and the secretion of prolactin, GH, LH, and TSH. *Annu Rev. Physiol.*, *41* 603, 1979

Labrie, F., *et al*: Mechanism of action of hypothalamic hormones in the adenohypophysis. *Annu Rev Physiol*, *41*:555, 1979.
Lewis, U. J.: Variants of growth hormone and prolactin and their posttranslational modifications. *Annu Rev Physiol*, *46*:33, 1984.
Li, C. H. (ed.): Hypothalamic Hormones. New York, Academic Press, 1979.
McCann, S. M., and Ojeda, S. R.: The role of brain monoamines, acetylcholine and prostaglandins in the control of anterior pituitary function. *In* DeGroot, L. J., *et al* (eds.): Endocrinology. Vol. 1. New York, Grune & Stratton, 1979, p. 55.
McKerns, K. W., and Jutisz, M. (eds.): Synthesis and Release of Adenohypophyseal Hormones. New York, Plenum Press, 1979.
Menninger, R. P.: Current concepts of volume receptor regulation of vasopressin release. *Fed Proc*, *44*:55, 1985.
Merimee, T. J.: Growth hormone: Secretion and action. *In* DeGroot, L. J., *et al* (eds.): Endocrinology. Vol. 1. New York, Grune & Stratton, 1979, p. 123.
Nissley, S. P., and Rechler, M. M.: Somatomedin insulin like growth factor tissue receptors. *Clin Endocrinol Metab*, *13*:43, 1984.
Ontjes, D. A., *et al*: The anterior pituitary gland. *In* Bondy, P. K., and Rosenberg, L. E. (eds.): Metabolic Control and Disease, 8th Ed. Philadelphia, W. B. Saunders Co., 1980, p. 1165.
Oriordan, J. L.: Recent Advances in Endocrinology and Metabolism. New York, Churchill Livingstone, 1982.
Pulleri, A., and MacLeod, R. M. (eds.): Neuroendocrinology: Biological and Clinical Aspects. New York, Academic Press, 1979.
Poste, G.: New insights into receptor regulation. *J Appl Physiol*, *57*:1297, 1984.
Rennels, E. G., and Herbert, D. C.: Functional correlates of anterior pituitary cytology. *In* Greep, R. O. (ed.): International Review of Physiology, Reproductive Physiology III. Vol. 22. Baltimore, University Park Press, 1980, p. 1.
Ryan, W. G., *et al* (eds.): Endocrine Disorders: A Pathophysiologic Approach. 2nd Ed. Chicago, Year Book Medical Publishers, 1980.
Sawyer, W. H., and Manning, M.: The use of antagonists of vasopressin in studies of its physiological functions. *Fed Proc*, *44*:78, 1985.
Schulster, D., and Levitski, A. (eds.): Cellular Receptors for Hormones and Neurotransmitters. New York, John Wiley & Sons, 1980.
Sklar, A. H., and Schrier, R. W.: Central nervous system mediators of vasopressin release. *Physiol Rev*, *63*:1243, 1983.
Sowers, J. R. (eds.): Hypothalamic Hormones. Stroudsburg, Pa., Dowden, Hutchinson & Ross, 1980.
Swanson, L. W., and Sawchenko, P. E.: Hypothalamic integration: Organization of the paraventricular and supraoptic nuclei. *Ann Rev Neurosci*, *6*:269, 1983.
Tepperman, J.: Metabolic and Endocrine Physiology: An Introductory Text. Chicago, Year Book Medical Publishers, 1980.
Tindall, G. T., and Collins, W. F. (eds.): Clinical Management: Pituitary Disorders. New York, Raven Press, 1979.
Verbert, H.: Oxytocin. *In* DeGroot, L. J., *et al* (eds.): Endocrinology. Vol. 1. New York, Grune & Stratton, 1979, p. 277.
Wass, J. A. H., and Besser, G. M.: The medical management of hormone secreting tumors of the pituitary. *Annu Rev Med*, *34*:283, 1983.
Weitzman, R., and Kleeman, C. R.: Water metabolism and the neurohypophyseal hormones. *In* Bondy, P. K., and Rosenberg, L. E. (eds.): Metabolic Control and Disease, 8th Ed. Philadelphia, W. B. Saunders Co., 1980, p. 1241.

76

The Thyroid Metabolic Hormones

The thyroid gland, which is located immediately below the larynx on either side of and anterior to the trachea, secretes two significant hormones, *thyroxine* and *triiodothyronine*, that have a profound effect on the metabolic rate of the body. It also secretes *calcitonin*, an important hormone for calcium metabolism that will be considered in detail in Chapter 79. Complete lack of thyroid secretion usually causes the basal metabolic rate to fall about 40 per cent below normal, and extreme excesses of thyroid secretion can cause the basal metabolic rate to rise as high as 60 to 100 per cent above normal. Thyroid secretion is controlled primarily by thyroid-stimulating hormone secreted by the anterior pituitary gland.

The purpose of this chapter is to discuss the formation and secretion of the thyroid hormones, their functions in the metabolic scheme of the body, and regulation of their secretion.

FORMATION AND SECRETION OF THE THYROID HORMONES

About 90 per cent of the hormone secreted by the thyroid gland is *thyroxine* and 10 per cent is *triiodothyronine*. However, a considerable portion of the thyroxine is converted to triiodothyronine in the blood and peripheral tissues, so that both are very important functionally. The functions of these two hormones are qualitatively the same, but they differ in rapidity and intensity of action. Triiodothyronine is about four times as potent as thyroxine, but it is present in the blood in much smaller quantities and persists for a much shorter time than does thyroxine.

Physiologic Anatomy of the Thyroid Gland. The thyroid gland is composed, as shown in Figure 76–1, of large numbers of closed *follicles* (150 to 300 microns in diameter) filled with a secretory substance called *colloid* and lined with *cuboidal epithelioid cells* that secrete into the interior of the follicles. The major constituent of colloid is the large glycoprotein *thyroglobulin*, which contains the thyroid hormones within its molecule. Once the secretion has entered the follicles, it must be absorbed back through the follicular epithelium into the blood before it can function in the body. The thyroid gland has a blood flow about five times the weight of the gland each minute, which is a blood supply as rich as that of any other area of the body with the probable exception of the adrenal cortex.

IODINE REQUIREMENTS FOR FORMATION OF THYROXINE

To form normal quantities of thyroxine, approximately 50 mg of ingested iodine are required *each year*, or approximately *1 mg per week*. To prevent iodine deficiency, common table salt is iodized with one part sodium iodide to every 100,000 parts sodium chloride.

Fate of Ingested Iodides. Iodides ingested orally are absorbed from the gastrointestinal tract into the blood in approximately the same manner as chlorides, but iodides do not remain in the circulatory system for a prolonged time because the kidneys have a very high plasma clearance for iodide ion—about 35 ml per minute in comparison with only 1 ml per minute for chloride ion. Within the first three days, four fifths of the ingested iodides are normally lost into the urine, and almost all the remaining one fifth is selectively removed from the circulating blood by the cells of the thyroid gland and used for synthesis of the thyroid hormones.

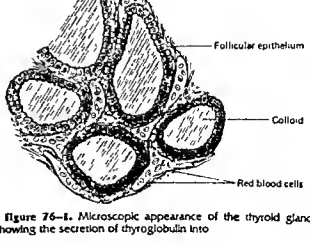


Figure 76–1. Microscopic appearance of the thyroid gland, showing the secretion of thyroglobulin into

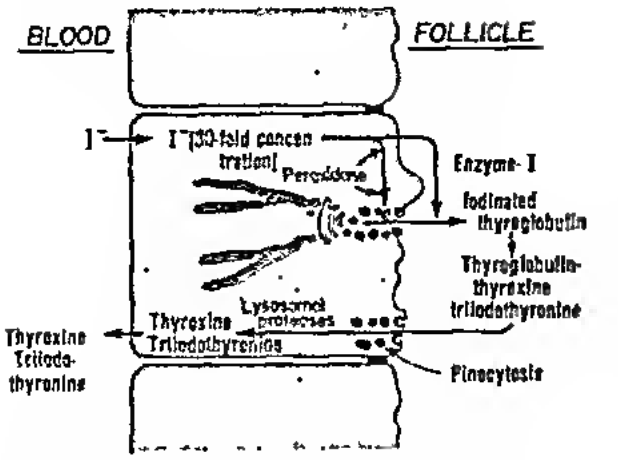


Figure 76–2. Thyroid cellular mechanisms for iodine transport, thyroxine (and triiodothyronine) formation, and thyroxine (and triiodothyronine) release into the blood.

THE IODIDE PUMP (IODIDE TRAPPING)

The first stage in the formation of thyroid hormones, as shown in Figure 76–2, is transport of iodides from the extracellular fluid into the thyroid glandular cells and follicles The basal membrane of the thyroid cell haa a specific ability to transport the iodide actively to the interior of the cell. This is called *iodide trapping*. In a normal gland, the iodide pump concentrates the iodide to about 30 times its concentration in the blood. However, when the thyroid gland becomes maximally active, the concentration ratio can rise to several times this value

THYROGLOBULIN AND CHEMISTRY OF THYROXINE AND TRIIODOTHYRONINE FORMATION

Formation and Secretion of Thyroglobulin by the Thyroid Cells. The thyroid cells are typical protein-secreting glandular cells as illustrated in Figure 76–2 The endoplasmic reticulum and Golgi apparatus synthesize and secrete into the follicles a large glycoprotein molecule called *thyroglobulin* with a molecular weight of 660,000

Each molecule of thyroglobulin contains 140 tyrosine amino acids, and these are the major substrates that combine with iodine to form the thyroid hormones These hormones form *within* the thyroglobulin molecule. That is, the tyrosine amino acid residues, as well as the thyroxine and triiodothyronine hormones formed from them, remain a part of the thyroglobulin molecule during the process of synthesis of the thyroid hormones

In addition to secreting the thyroglobulin the glandular cells also provide the iodine, the enzymes, and other substances necessary for thyroid hormone synthesis.

Oxidation of the Iodide Ion. The first essential step in the formation of the thyroid hormones is conversion of the iodide ions to an *oxidized form of iodine* that is then capable of combining directly with the amino acid tyrosine.

This oxidation of iodine is promoted by the enzyme *peroxidase* and its accompanying *hydrogen peroxide*, which provide a potent system capable of oxidizing iodides. The peroxidase is located either in the apical membrane of the cell or in the cytoplasm immediately adjacent to this membrane, thus providing the oxidized iodine at exactly the point in the cell where the thyroglobulin molecule first issues forth from the Golgi apparatus. When the peroxidase system is blocked, or when it is hereditarily absent from the cells, the rate of formation of thyroid hormones falls to zero.

Iodination of Tyrosine and Formation of the Thyroid Hormones—"Organification" of Thyroglobulin. The binding of iodine with the thyroglobulin molecule is called *organification* of the thyroglobulin. Oxidized iodine even in the molecular form will bind directly but slowly with the amino acid tyrosine, but in the thyroid cells the oxidized iodine is associated with an *iodinase* enzyme that causes the process to occur within seconds or minutes Therefore, almost as rapidly as the thyroglobulin molecule is released from the Golgi apparatus, or as it is secreted through the apical cell membrane into the follicle, iodine binds with about one sixth of the tyrosine residues within the thyroglobulin molecule.

Figure 76–3 illustrates the successive stages of iodination of tyrosine and the final formation of the two important thyroid hormones, thyroxine and triiodothyronine. Tyrosine is first iodized to *monoiodotyrosine* and then to *diiodotyrosine*. Then during the next few minutes, hours, and even days, more and more of the diiodotyrosine residues become *coupled* with each other (though the mechanism of the coupling is not understood). The product of the coupling reaction is the molecule *thyroxine* that also remains part of the thyroglobulin molecule Or one molecule of monoiodotyrosine couples with one molecule of diiodotyrosine to form *triiodothyronine*.

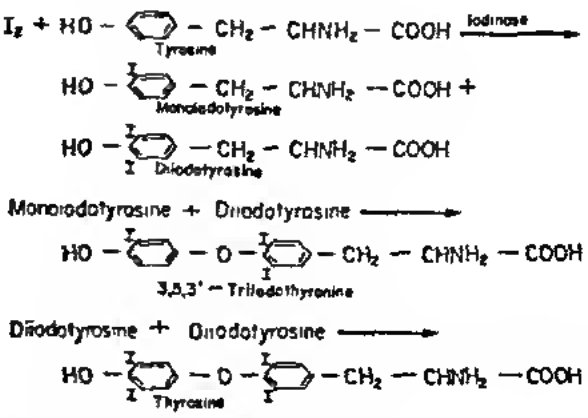


Figure 76–3. Chemistry of thyroxine and triiodothyronine formation.

Storage of Thyroglobulin. After synthesis of the thyroid hormones has run its course, each thyroglobulin molecule contains from 5 to 6 thyroxine molecules, and there is also an average of about 1 triiodothyronine molecule for every 10 molecules of thyroxine. In this form the thyroid hormones are often stored in the follicles for several months. In fact, the total amount stored is sufficient to supply the body with its normal requirements of thyroid hormones for one to three months. Therefore, even when synthesis of thyroid hormone ceases entirely, the effects of deficiency might not be observed for several months

RELEASE OF THYROXINE AND TRIIODOTHYRONINE FROM THYROGLOBULIN

Thyroglobulin itself is not released into the circulating blood in significant amounts; instead, the thyroxine and triiodothyronine are first cleaved from the thyroglobulin molecule, and then these free hormones are released This process occurs as follows: The apical surface of the thyroid cells normally sends out pseudopod extensions that close around small portions of the colloid to form pinocytic vesicles. Then lysosomes immediately fuse with these vesicles to form digestive vesicles containing the digestive enzymes from the lysosomes mixed with the colloid The *proteinases* among these enzymes digest the thyroglobulin molecules and release the thyroxine and triiodothyronine, which then diffuse through the base of the thyroid cell into the surrounding capillaries. Thus, the thyroid hormones are released into the blood.

About three quarters of the iodinated tyrosine in the thyroglobulin never becomes thyroid hormones but instead remains monoiodotyrosine or diiodotyrosine. During the digestion of the thyroglobulin molecule to cause release of thyroxine and triiodothyronine, these iodinated tyrosines also are freed from the thyroid cells. However, they are not secreted into the blood. Instead, their iodine is cleaved from them by a *deiodinase enzyme* that makes most of this iodine available for recycling within the gland for forming additional thyroid hormones. In the congenital absence of this deiodinase enzyme, persons frequently become iodine-deficient because of failure of this recycling process.

Daily Rate of Secretion of Thyroxine and Triiodothyronine. Over 90 per cent of the thyroid hormone released from the thyroid gland is thyroxine, and less than 10 per cent is triiodothyronine. However, during the ensuing few days while these hormones circulate in the blood, small portions of the thyroxine are slowly deiodinated to form additional triiodothyronine. Therefore, the quantities of the two hormones finally delivered to the tissues is approximately 90 micrograms of thyroxine per day and 35 micrograms of triiodothyronine per day. (Another 35 micrograms of so-called "reverse triiodothyronine" are formed each day by removal of one of the iodines of thyroxine near its carboxyl end. However, reverse triiodothyronine is almost totally inactive.)

Once the two hormones finally enter the peripheral tissue cells, triiodothyronine is about four times as potent in stimulating metabolism and causing other intracellular effects as thyroxine. On the other hand, the duration of action of thyroxine is four or more times as long as the duration of action of triiodothyronine. Therefore, the integrated effect of each of the hormones over the period of its action per unit mass of delivered hormone is probably about equal.

On entering the peripheral cells, most of the thyroxine is deiodinated to form triiodothyronine Therefore, it is believed that the true intracellular hormone is principally triiodothyronine rather than thyroxine.

TRANSPORT OF THYROXINE AND TRIIODOTHYRONINE TO THE TISSUES

Binding of Thyroxine and Triiodothyronine with Plasma Proteins. On entering the blood, all but minute portions of the thyroxine and triiodothyronine combine immediately with several of the plasma proteins. They combine approximately as follows: two thirds with *thyroxine-binding globulin*, which is a glycoprotein; about one fourth with *thyroxine-binding prealbumin*; and about one tenth with *albumin*. The quantity of thyroxine-binding globulin in the blood is only 1 to 1.5 milligrams per deciliter of plasma, but its affinity for the thyroid hormones is so great that it still binds most of the hormones. Its affinity (and that of the other plasma proteins) is over ten times as great for thyroxine as for triiodothyronine. This difference, plus the fact that the concentration of thyroxine in the plasma is considerably greater than that of triiodothyronine, causes the total amount of protein-bound thyroxine to be about 60 times as great as the protein-bound triiodothyronine.

Release of Thyroxine and Triiodothyronine to the Tissue Cells. Because of the very high affinity of the plasma-binding proteins for the thyroid hormones, these substances—in particular, thyroxine—are released to the tissue cells ooly very slowly. Half of the thyroxine in the blood is released to the tissue cells approximately every 6 days, whereas half of the triiodothyronine—because of its lower affinity—is released to the cells in approximately 1.3 days.

On eoteriog the cells, both these hormones again biod with intracellular proteins, the thyroxine ooce again binding more strongly than the triiodothyronine. Therefore, they are agaio stored, but

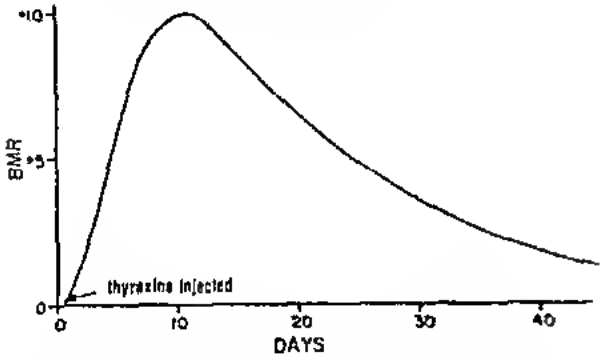


Figure 76–4. Prolonged effect on the basal metabolic rate caused by administering a single large dose of thyroxine

this time in the functional cells themselves, and they are used slowly over a period of days or weeks.

Latency and Duration of Action of the Thyroid Hormones. After injection of a large quantity of thyroxine into a human being, essentially no effect on the metabolic rate can be discerned for two to three days, thereby illustrating that there is a *long latent* period before thyroxine activity begins. Once activity does begin, it increases progressively and reaches a maximum in 10 to 12 days, as shown in Figure 76–4. Thereafter, it decreases with a half-life of about 15 days. Some of the activity still persists as long as 6 weeks to 2 months later.

The actions of triiodothyronine occur about four times as rapidly as those of thyroxine, with the latent period as short as 6 to 12 hours and maximum cellular activity occurring within 2 to 3 days.

A large share of the latency and prolonged period of action of these hormones is caused by their binding with proteins both in the plasma and in the tissue cells, followed by their slow release. However, we shall see in subsequent discussions that part of the latent period also results from the manner in which these hormones perform their functions in the cells themselves.

FUNCTIONS OF THE THYROID HORMONES IN THE TISSUES

The thyroid hormones have two major effects on the body: (1) an increase in the overall metabolic rate, and (2) in children, stimulation of growth.

GENERAL INCREASE IN METABOLIC RATE

The thyroid hormones increase the metabolic activities of almost all tissues of the body (with a few notable exceptions such as the brain, retina, spleen, testes, and lungs). The basal metabolic rate can increase to as much as 60 to 100 per cent above normal when large quantities of the hormones are secreted. The rate of utilization of foods for energy is greatly accelerated. The rate of protein synthesis is at times increased, while at the same time the rate of protein catabolism is also increased. The growth rate of young persons is greatly accelerated. The mental processes are excited, and the activity of many other endocrine glands is often increased. Yet despite the fact that we know all these many changes in metabolism under the influence of the thyroid hormones, the basic mechanism (or mechanisms) by which the hormones function is much less well known. However, some of the mechanisms of action of the thyroid hormones are described in the following sections.

Effect of Thyroid Hormones on Causing Increased Protein Synthesis. When either thyroxine or triiodothyronine is given to an animal, protein synthesis increases in almost all tissues of the body. The first stage of the increased protein synthesis begins almost immediately and results from stimulation of the *translation* process—that is, increase in the rate of formation of proteins by the ribosomes. The second stage occurs hours to days later and is caused by increased RNA synthesis by the genes, the process of *transcription*, which leads to a generalized increase in synthesis of many types of proteins within the cells. This stimulation of the genes occurs in the following way: (1) The thyroid hormone, probably mainly triiodothyonine derived from deiodinated thyroxine, within the cell, combines with a "receptor" protein in the cell nucleus. (2) This combination, or the product of it, then activates a large portion of the cellular genes to cause RNA formation and subsequent protein formation.

Effect of Thyroid Hormones on the Cellular Enzyme Systems. Within a week or so following administration of the thyroid hormones, at least 100 and probably many more intracellular enzymes are increased in quantity because of the increased protein synthesis. As an example, one enzyme, α-glycerophosphate dehydrogenase, can be increased to an activity six times its normal level. Since this enzyme is particularly important in the degradation of carbohydrates, its increase could help explain the rapid utilization of carbohydrates under the influence of thyroxine. Also the oxidative enzymes and the elements of the intracellular electron transport system, both of which are found in mitochondria, are greatly increased.

Effect of Thyroid Hormones on Mitochondria. When thyroxine or triiodothyronine is given to an animal, the mitochondria in most cells of the body increase in size and also in number. Furthermore, the total membrane surface area of the mitochondria increases almost directly in proportion to the increased metabolic rate of the whole animal. Therefore, it seems almost to be an obvious deduction that the principal function of thyroxine might be simply to increase the number and activity of mitochondria, and these in turn increase the

rate of formation of ATP to energize cellular function. Unfortunately, though, the increase in number and activity of mitochondria could as well be the *result* of increased activity of the cells as be the cause of the increase.

When *extremely* high concentrations of thyroid hormone are administered, the mitochondria swell inordinately, and there is uncoupling of the oxidative phosphorylation process with production of large amounts of heat but little ATP However, under natural conditions, it is questionable whether the concentration of thyroid hormones becomes high enough to cause this effect even in persons who have thyrotoxicosis.

Effect of Thyroid Hormone in Increasing Active Transport of Ions Through Cell Membranes. One of the enzymes that becomes increased in response to thyroid hormone is *Na-K ATPase*. This in turn increases the rate of transport of both sodium and potassium through the cell membranes of some tissues. Since this process utilizes energy and also increases the amount of heat produced in the body, it has also been suggested that this might be one of the mechanisms by which thyroid hormone increases the body's metabolic rate.

Summary. It is clear that we know many effects that occur in the cells throughout the body under the influence of thyroid hormone. Yet, a specific metabolic mechanism that leads to all of these effects has been elusive At present, the most likely basic function of the thyroid hormones is their capability to activate the DNA transcription process in the cell nucleus with resulting formation of many new cellular proteins.

EFFECT OF THYROID HORMONE ON GROWTH

Thyroid hormone has both general and specific effects on growth. For instance, it has long been known that thyroid hormone is essential for the metamorphic change of the tadpole into the frog. In the human being, the effect of thyroid hormone on growth is manifest mainly in growing children. In those who are hypothyroid, the rate of growth is greatly retarded. In those who are hyperthyroid, excessive skeletal growth often occurs, causing the child to become considerably taller than otherwise. However, the epiphyses close at an early age so that the duration of growth, and the eventual height of the adult, may be shortened.

An important effect of thyroid hormone is to promote growth and development of the brain during fetal life and for the first few years of postnatal life. If the fetus does not secrete sufficient quantities of thyroid hormone, growth and maturation of the brain both before birth and afterward are greatly retarded. Without specific thyroid therapy within days or weeks after birth, the child will remain mentally deficient throughout life. This will be discussed more fully later in the chapter.

The growth-promoting effect of thyroid hormone is presumably based on its ability to promote protein synthesis. On the other hand, a great excess of thyroid hormone can cause more rapid catabolism than synthesis of protein, so that the protein stores are then actually mobilized and amino acids released into the extracellular fluids.

EFFECTS OF THYROID HORMONE ON SPECIFIC BODILY MECHANISMS

Effect on Carbohydrate Metabolism. Thyroid hormone stimulates almost all aspects of carbohydrate metabolism, including rapid uptake of glucose by the cells, enhanced glycolysis, enhanced gluconeogenesis, increased rate of absorption from the gastrointestinal tract, and even increased insulin secretion with its resultant secondary effects on carbohydrate metabolism. All these effects probably result from the overall increase in enzymes caused by thyroid hormone

Effect on Fat Metabolism. Essentially all aspects of fat metabolism are also enhanced under the influence of thyroid hormone. However, since fats are the major source of long-term energy supplies, the fat stores of the body are depleted to a greater extent than are most of the other tissue elements; in particular, lipids are mobilized from the fat tissue, which increases the free fatty acid concentration in the plasma, and thyroid hormona also greatly accelerates the oxidation of free fatty acids by the cells.

Effect on Blood and Liver Fats. Increased thyroid hormone *decreases* the quantity of cholesterol, phospholipids, and triglycerides in the blood, even though it *increases* the free fatty acids, On the other hand, decreased thyroid secretion greatly increases the concentrations of cholesterol, phospholipids, and triglycerides and almost always causes excessive deposition of fat in the liver. The large increase in circulating blood lipids in prolonged hypothyroidism is often associated with severe arteriosclerosis, which was discussed in Chapter 68.

Effect on Vitamin Metabolism. Because thyroid hormone increases the quantities of many of the different enzymes and because vitamins are essential parts of some of the enzymes or coenzymes, thyroid hormone causes increased need for vitamins. Therefore, a relative vitamin deficiency can occur when excess thyroid hormone is secreted, unless at the same time increased quantities of vitamins are available.

Effect on Basal Metabolic Rate. Because thyroid hormone increases metabolism in most cells of the body (with the exception of the brain, retina, spleen, testes, and lungs), excessive quantities of the hormone can occasionally increase the basal metabolic rate to as much as 60 to 100 per cent above normal. On the other hand, when no thyroid hormone is produced, the basal metabolic rate falls almost to half normal; that is, the basal metabolic rate becomes −30 to −45, as discussed in Chapter 71. Figure 76–5 shows the approximate relationship between the daily supply of thyroid hormones and the basal metabolic rate. Extreme amounts of the hormones are required to cause very high basal metabolic rates

Effect on Body Weight. Greatly increased thyroid

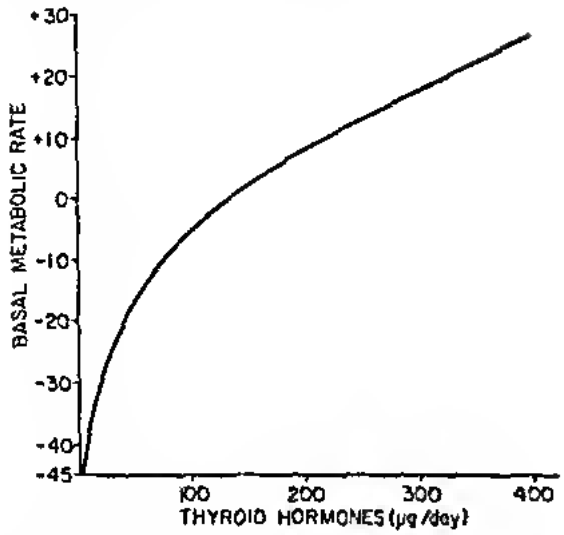


Figure 76–5. Approximate relationship of thyroid hormone daily rate of secretion to the basal metabolic rate

hormone production almost always decreases the body weight, and greatly decreased production almost always increases the body weight, but these effects do not always occur, because thyroid hormone increases the appetite, and this may overbalance the change in the metabolic rate

Effect on the Cardiovascular System. ***Blood Flow and Cardiac Output.*** Increased metabolism in the tissues causes more rapid utilization of oxygen than normally and causes greater than normal quantities of metabolic end-products to be released from the tissues These effects cause vasodilatation in most of the body tissues, thus increasing blood flow Especially does the rate of blood flow in the skin increase because of the increased necessity for heat elimination

As a consequence of the increased blood flow, the cardiac output also increases, sometimes rising to 50 per cent or more above normal when excessive thyroid hormone is present.

Heart Rate. The heart rate increases considerably more under the influence of thyroid hormone than would be expected simply because of the increased cardiac output Therefore, thyroid hormone probably has a direct effect on the excitability of the heart, which in turn increases the heart rate This effect is of particular importance because the heart rate is one of the most sensitive physical signs that the clinician has for determining whether a patient has excessive or diminished thyroid hormone production

Strength of Heartbeat. The increased enzymatic activity caused by increased thyroid hormone production apparently increases the strength of the heart when only a slight excess of thyroid hormone is secreted. This is analogous to the increase in strength of heart beat that occurs in mild fevers and during exercise However, when thyroid hormone is increased markedly, the heart muscle strength becomes depressed because of excessive protein catabolism Indeed, some severely thyrotoxic patients die of cardiac decompensation secondary to myocardial failure and increased cardiac load imposed by the increased output

Blood Volume. Thyroid hormone causes the blood volume to increase slightly. The effect probably results least partly from the vasodilatation which allows increased quantities of blood to collect in the circulatory system

Arterial Pressure. The *mean* arterial pressure usually is unchanged. However, because of the increased stroke volume with each heart beat and increased blood flow through the tissues between heart beats, the pulse pressure is increased, with the systolic pressure elevated 10 to 20 mm Hg, and the diastolic pressure correspondingly reduced.

Effect on Respiration. The increased rate of metabolism increases the utilization of oxygen and the formation of carbon dioxide, these effects activate all the mechanisms that increase the rate and depth of respiration

Effect on the Gastrointestinal Tract. In addition to increased appetite and food intake, which has been discussed, thyroid hormone increases both the rate of secretion of the digestive juices and the motility of the gastrointestinal tract. Often, diarrhea results Lack of thyroid hormone causes constipation

Effect on the Central Nervous System. In general thyroid hormone increases the rapidity of cerebration but also often dissociates this, while, on the other hand lack of thyroid hormone decreases this function The hyperthyroid individual is likely to develop extreme nervousness and is likely to have many psychoneurotic tendencies, such as anxiety complexes, extreme worry, or paranoias

Effect on the Function of the Muscles. Slight increase in thyroid hormone usually makes the muscles react with vigor, but when the quantity of hormone becomes excessive, the muscles become weakened because of excess protein catabolism. On the other hand, lack of thyroid hormone causes the muscles to become extremely sluggish, and they relax slowly after a contraction

Muscle Tremor. One of the most characteristic signs of hyperthyroidism is a fine muscle tremor. This is not the coarse tremor that occurs in Parkinson's disease or in shivering, for it occurs at the rapid frequency of 10 to 15 times per second The tremor can be observed easily by placing a sheet of paper on the extended fingers and noting the degree of vibration of the paper. This tremor is believed to be caused by increased sensitivity of the neuronal synapses in the areas of the cord that control muscle tone The tremor is an important means for assessing the degree of thyroid hormone effect on the central nervous system

Effect on Sleep. Because of the exhausting effect of thyroid hormone on the musculature and on the central nervous system, the hyperthyroid subject often has a feeling of constant tiredness, but because of the excitable effects of thyroid hormone on the synapses, it is difficult to sleep On the other hand, extreme somnolence is characteristic of hypothyroidism.

Effect on Other Endocrine Glands. Increased thyroid hormone increases the rates of secretion of most other endocrine glands, but it also increases the need of the tissues for the hormones For instance, increased thyroxine secretion increases the rate of glucose metabolism everywhere in the body and therefore causes a corresponding need for increased insulin secretion by the pancreas. Also, thyroid hormone increases many metabolic activities related to bone formation and, as a consequence, increases the need for parathyroid hormone. And, finally, thyroid hormone increases the rate at which adrenal glucocorticoids are inactivated by the

liver. This leads to feedback increase in ACTH production by the anterior pituitary and, therefore, increased rate of glucocorticoid secretion by the adrenal glands also.

Effect of Thyroid Hormone on Sexual Function. For normal sexual function to occur, thyroid secretion needs to be approximately normal—neither too great nor too little. In men, lack of thyroid hormone is likely to cause complete loss of libido, while, on the other hand, great excesses of the hormone frequently cause impotence. In women, lack of thyroid hormone often causes *menorrhagia* and *polymenorrhea*, which mean, respectively, excessive and frequent menstrual bleeding. In some women the lack may cause irregular periods and, occasionally, even total *amenorrhea*. A hypothyroid woman, like a man, is also likely to have greatly decreased libido. Conversely, in the hyperthyroid woman, *oligomenorrhea*, which means greatly reduced bleeding, is usual, and occasionally amenorrhea results.

The action of thyroid hormone on the gonads cannot be pinpointed to a specific function but probably results from a combination of direct metabolic effects on the gonads and of excitatory and inhibitory effects operating through the anterior pituitary hormones that control the sexual functions.

REGULATION OF THYROID HORMONE SECRETION

To maintain normal levels of metabolic activity in the body, precisely the right amount of thyroid hormone must be secreted all the time, and to provide this, specific feedback mechanisms operate through the hypothalamus and anterior pituitary gland to control the rate of thyroid secretion. These mechanisms can be explained as follows:

Effects of Thyroid-Stimulating Hormone on Thyroid Secretion. Thyroid-stimulating hormone (TSH), also known as *thyrotropin*, is an anterior pituitary hormone, a glycoprotein with a molecular weight of about 28,000, that was discussed in Chapter 74; it increases the secretion of thyroxine and triiodothyronine by the thyroid gland. Its specific effects on the thyroid gland are

(1) increased proteolysis of the thyroglobulin that has already been stored in the follicles, with resultant release of the thyroid hormones into the circulating blood and diminishment of the follicular substance itself;

(2) increased activity of the iodide pump, which increases the rate of "iodide trapping" in the glandular cells, sometimes increasing the ratio of intracellular to extracellular iodide concentration severalfold;

(3) increased iodination of tyrosine and increased coupling to form the thyroid hormones;

(4) increased size and increased secretory activity of the thyroid cells; and

(5) increased number of thyroid cells, plus a change from cuboidal to columnar cells and much infolding of the thyroid epithelium into the follicles.

In summary, thyroid-stimulating hormone *increases all the known activities of the thyroid glandular cells.*

The most important early effect following administration of thyroid-stimulating hormone is proteolysis of the thyroglobulin, which causes release of thyroxine and triiodothyronine into the blood within 30 minutes. The other effects require hours or even days and weeks to develop fully.

Role of Cyclic AMP in the Stimulntory Effect of TSH. In the past it was difficult to explain the many and varied effects of thyroid-stimulating hormone on the thyroid cell. However, it is now clear that at least most of these result from activation of the "second messenger" *cyclic AMP* system of the cell. The first event in this activation is binding of the thyroid-stimulating hormone with specific TSH receptors on the basal membrane surfaces of the cell. This then activates *adenylcyclase* in the membrane, which increases the formation of cyclic AMP in the cell. Finally, the cyclic AMP acts as a *second messenger* to activate essentially all enzyme systems of the thyroid cell. The result is both an immediate increase in secretion of thyroid hormones and prolonged growth of the thyroid glandular tissue itself. This method for control of thyroid cell activity is similar to the function of cyclic AMP in many other target tissues of the body.

Hypothalamic Regulation of TSH Secretion by the Anterior Pituitary—Thyrotropin-Releasing Hormone (TRH)

Electrical stimulation of multiple areas of the hypothalamus increases the anterior pituitary secretion of TSH and correspondingly increases the activity of the thyroid gland. This control of anterior pituitary secretion is exerted by a hypothalamic hormone, *thyrotropin-releasing hormone* TRH), which is secreted by nerve endings in the median *eminence of the hypothalamus* and then transported from there to the anterior pituitary in the hypothalamic-hypophysial portal blood, as was explained in Chapter 74. The precise nuclei of the hypothalamus that are responsible for causing secreting of TRH in the median eminence are not known. However, injection of radioactive antibodies that attach specifically to TRH have shown this hormone to be present in many different hypothalamic loci, including the (1) *dorsomedial nucleus*, (2) *suprachiasmatic nucleus*, (3) *ventromedial nucleus*, (4) *anterior hypothalamus*, (5) *preoptic area*, and (6) *paraventricular nucleus*.

TRH has been obtained in pure form, and it has proved to be a very simple substance, a tripeptide amide—*pyroglutamyl-histidyl-proline-amide*. TRH directly affects the anterior pituitary gland cells to increase their output of thyroid-stimulating hormone. When the portal system from the hypothalamus to the anterior pituitary gland is completely

blocked, the rate of secretion of TSH by the anterior pituitary is greatly decreased but not reduced to zero.

The hypothalamus can also inhibit anterior pituitary secretion of thyroid-stimulating hormone. It does this by secreting *somatostatin*, which inhibits the secretion of TSH at the same time that it also inhibits growth hormone secretion. But the role of somatostatin in the overall thyroid control system is unknown.

Effects of Cold and Other Neurogenic Stimuli on TRH and TSH Secretion. One of the best-known stimuli for increasing the rate of TRH secretion by the hypothalamus, and therefore TSH secretion by the anterior pituitary, is exposure of an animal to cold. This effect almost certainly results from excitation of the anterior area of the hypothalamus by the cold stimulus. Exposure of rats for several weeks to very severe cold increases the output of thyroid hormones sometimes more than 100 per cent and can increase the basal metabolic rate as much as 50 per cent. Indeed, people moving to arctic regions have been known to develop basal metabolic rates 15 to 20 per cent above normal; however, the behavioral propensity of human beings to protect themselves from cold usually prevents a measurable effect.

Various emotional reactions can also affect the output of TRH and TSH and can, therefore, indirectly affect the secretion of thyroid hormones. On the other hand, excitement and anxiety—conditions that greatly stimulate the sympathetic nervous system—cause acute decrease in secretion of TSH, perhaps because these states increase the metabolic rate and the body heat.

Neither these emotional effects nor the effect of cold is observed after the hypophysial stalk has been cut, illustrating that both these effects are mediated by way of the hypothalamus.

Inverse Feedback Effect of Thyroid Hormone on Anterior Pituitary Secretion of TSH—Feedback Regulation of Thyroid Secretion

Increased thyroid hormone in the body fluids decreases the secretion of TSH by the anterior pituitary. When the rate of thyroid hormone secretion rises to about 1.75 times normal, the rate of TSH secretion falls essentially to zero. Most of this feedback depressant effect occurs even when the anterior pituitary has been completely separated from the hypothalamus, but the effect is somewhat greater if the hypothalamus and hypothalamic-hypophysial portal system are intact. Therefore, as illustrated in Figure 76–6, it is probable that increased thyroid hormone inhibits anterior pituitary secretion of TSH mainly by a direct feedback effect on the anterior pituitary itself but perhaps secondarily by much weaker effects acting through the hypothalamus.

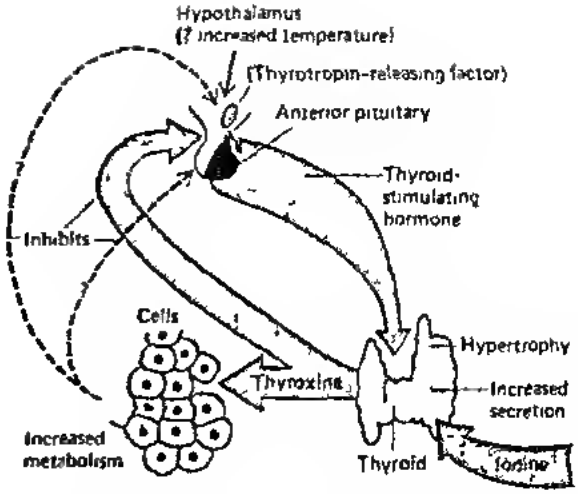


Figure 76–6. Regulation of thyroid secretion.

One mechanism that has been suggested for the feedback effect on the anterior pituitary gland is that thyroid hormone reduces the number of TRH receptors on the cells that secrete thyroid-stimulating hormone. Therefore, the stimulating effect on these cells by TRH from the hypothalamus is greatly reduced.

Regardless of the mechanism of the feedback, its effect is to maintain an almost constant concentration of free thyroid hormone in the circulating body fluids. For instance, during periods of heavy exercise, thyroid hormone is consumed much more rapidly than under resting conditions; yet, because of appropriate feedback control, the rate of secretion of thyroid hormone rises to equal the rate of consumption, and the blood thyroid hormone concentration remains almost exactly constant.

If there is a feedback effect through the hypothalamus in addition to the direct feedback to the pituitary gland itself, this probably operates very slowly and might be caused at least partly by changes in the rate of metabolism in the body temperature controlling centers of the hypothalamus, which are known to have a significant effect in controlling the thyroid hormone system.

ANTITHYROID SUBSTANCES

Drugs that suppress thyroid secretion are called antithyroid substances. The three best known of these are thiocyanate, propylthiouracil, and high concentrations of inorganic iodides. The mechanism by which each of these blocks thyroid secretion is different from the others, and they can be explained as follows:

Decreased Iodide Trapping Caused by Thiocyanate Ions. The same active pump that pumps iodide ions into the thyroid cells can also pump thiocyanate ions, perchlorate ions, and nitrate ions. Therefore, administration of thiocyanate (or one of the other ions as well) in high enough concentration can cause competitive inhibition of iodide transport into the cell—that is, inhibition of the iodide-trapping mechanism.

The decreased availability of iodide in the glandular cells does not stop the formation of thyroglobulin; it

merely causes thyroglobulin to be formed without becoming iodinated and therefore without forming the thyroid hormones. This in turn leads to increased secretion of TSH by the anterior pituitary gland, which then causes overgrowth of the thyroid gland even though this gland still does not secrete adequate quantities of thyroid hormones. Therefore, use of thiocyanates and some other ions to block thyroid secretion or the presence of these ions in certain foods can lead to the development of an enlarged thyroid gland, which is called a *goiter*

Depression of Thyroid Hormone Formation by Propylthiouracil. Propylthiouracil (and other similar compounds such as methimazole and carbimazole) prevents formation of thyroid hormone from iodides and tyrosine. The mechanism of this is partly to block iodination of tyrosine but even more so to block the coupling of two iodinated tyrosines to form thyroxine or triiodothyronine.

Propylthiouracil does not prevent formation of thyroglobulin, but the absence of thyroxine and triiodothyronine in the thyroglobulin leads to tremendous feedback enhancement of TSH secretion by the anterior pituitary gland. Therefore, in the same manner that thiocyanates cause the thyroid gland to enlarge, so also does propylthiouracil lead to enhanced growth of the glandular tissue, thus forming a goiter

Decrease in Thyroid Activity Caused by Iodides. When iodides are present in the blood *in high concentration* (100 times the normal plasma level), most activities of the thyroid gland are decreased, but often they remain decreased only for a few weeks. The rate of iodide trapping is reduced, the rate of thyroid hormone formation is decreased, the secretory activity of the thyroid cells is decreased, and the rate of thyroid hormone release from the thyroglobulin is decreased Since these are almost exactly opposite to the effects of TSH on the thyroid gland, it has been suggested that high concentrations of iodides in the blood directly inhibit the thyroid-stimulating effect of TSH.

Because iodides in high concentrations decrease all phases of thyroid activity, they decrease the size of the thyroid gland and especially decrease its blood supply, in contradistinction to the opposite effects caused by most of the other antithyroid agents. For this reason, iodides are frequently administered to patients for two or three weeks prior to surgical removal of the thyroid gland to decrease the necessary amount of surgery

DISEASES OF THE THYROID

HYPERTHYROIDISM

Most effects of hyperthyroidism are obvious from the preceding discussion of the various physiological effects of thyroid hormone. However, some specific effects should be mentioned in connection especially with the development, diagnosis, and treatment of hyperthyroidism.

Causes of Hyperthyroidism (Toxic Goiter, Thyrotoxicosis, Graves' Disease). In the patient with hyperthyroidism the entire thyroid gland is usually markedly hyperplastic. It is increased to two to three times normal size, with tremendous folding of the follicular cell lining into the follicles so that the number of cells is increased several times as much as the size of the gland is increased. Also, each cell increases its rate of secretion severalfold; radioactive iodine uptake studies indicate that these hyperplastic glands secrete thyroid hormone at a rate as great as 5 to 15 times normal.

These changes in the thyroid gland are similar to those caused by excessive thyroid-stimulating hormone. However, radioimmunoassay studies have shown the plasma TSH concentrations to be less than normal rather than enhanced in most of the patients, and often to be essentially zero On the other hand, some other substance or substances that have actions similar to that of TSH are found in the blood of almost all these patients. These substances are usually immunoglobulin antibodies that bind with the thyroid cell membranes. It is believed that they bind with the same membrane receptors that bind TSH and that this induces continual activation of the cyclic AMP system of the cells, with the resultant development of hyperthyroidism. One of these antibodies, found in 50 to 80 per cent of thyrotoxic patients, is called *long-acting thyroid stimulator* (LATS) This has a prolonged stimulating effect on the thyroid gland, lasting as long as 12 hours in contrast to a little over 1 hour for TSH The high level of thyroid hormone secretion caused by LATS, in turn, suppresses anterior pituitary formation of TSH

The antibodies that cause hyperthyroidism almost certainly develop as the result of autoimmunity that has developed against thyroid tissue Presumably, at some time in the history of the person an excess of thyroid cell antigens has been released from the thyroid cells, and this has resulted in the formation of antibodies against the thyroid gland itself

Aside from the aforementioned variety of hyperthyroidism, the condition occasionally also results from a localized adenoma (a tumor) that develops in the thyroid tissue and secretes large quantities of thyroid hormone This is different from the more usual type of hyperthyroidism in that it usually is not associated with evidence of autoimmune disease An interesting effect of the adenoma is that as long as it continues to secrete large quantities of thyroid hormone, function in the remainder of the thyroid gland is almost totally inhibited because the thyroid hormone from the adenoma depresses the production of TSH by the pituitary gland.

Symptoms of Hyperthyroidism

The symptoms of hyperthyroidism are obvious from the preceding discussion of the physiology of the thyroid hormones—intolerance to heat, increased sweating, mild to extreme weight loss (sometimes as much as 100 pounds), varying degrees of diarrhea, muscular weakness, nervousness or other psychic disorders, extreme fatigue but inability to sleep, and tremor of the hands.

Exophthalmos. Most, but not all, persons with hyperthyroidism develop some degree of protrusion of the eyeballs, as illustrated in Figure 76–7. This condition is called *exophthalmos*. A major degree of exophthalmos occurs in about one third of the hyperthyroid patients, and the condition on rare occasions becomes severe enough that the eyeball protrusion stretches the optic nerve enough to damage vision. Much more often, the eyes are damaged because the eyelids do not close completely when the person blinks or is asleep. As a result, the dry epithelial surfaces of the eyes become irritated and often infected, resulting in ulceration of the cornea.

The cause of the protruding eyes is edematous swell-

Figure 76–7. Patient with exophthalmic hyperthyroidism. Note protrusion of the eyes and retraction of the superior eyelids. The basal metabolic rate was +40 (Courtesy of Dr Leonard Posey)

ing of the retro-orbital tissues and degenerative changes in the extraocular muscles The factor or factors that initiate these changes are still in serious dispute. In most patients, antibodies can be found in the blood that react with the retro-orbital tissues Therefore, there is much reason to believe that exophthalmos, like hyperthyroidism itself, is an autoimmune process Usually, the exophthalmos disappears or at least greatly ameliorates with treatment of the hyperthyroidism

Diagnostic Tests for Hyperthyroidism. For the usual case of hyperthyroidism, the most accurate diagnostic test is direct measurement of the concentration of "free" thyroxine in the plasma, using appropriate radioimmunoassay procedures

Other tests that are sometimes used are

(1) The basal metabolic rate is usually increased to +30 or +60 in severe hyperthyroidism

(2) The stimulating effect of thyrotropin-releasing hormone (TRH) on the anterior pituitary gland is also often measured To do this, a small amount of synthetic thyrotropin-releasing hormone is injected intravenously, and the increase in concentration of thyroid-stimulating hormone (TSH) in the plasma is measured by radioimmunoassay In the usual type of thyrotoxicosis, the anterior pituitary secretion of TSH is so completely suppressed by the large amounts of circulating thyroxine and triiodothyronine that there will be almost no increase in plasma TSH

Physiology of Treatment in Hyperthyroidism. The most direct treatment for hyperthyroidism is surgical removal of most of the thyroid gland In general, it is desirable to prepare the patient for surgical removal of the gland prior to the operation. This is done by administering propylthiouracil, usually for several weeks, until the basal metabolic rate of the patient has returned to normal Then administration of high concentrations of iodides for two weeks immediately prior to operation causes the gland itself to recede in size and its blood supply to diminish. By using these preoperative procedures, the operative mortality is less than 1 in 1000 in the better hospitals, whereas prior to development of these procedures the operative mortality was as great as 1 in 25.

Treatment of the Hyperplastic Thyroid Gland with Radioactive Iodine. As much as 80 to 90 per cent of an injected dose of iodide is absorbed by the hyperplastic, toxic thyroid gland within a day after injectio If this injected iodine is radioactive, it can destr internally the secretory cells of the thyroid gland Us ally 5 millicuries of radioactive iodine is given to t patient, whose condition is reassessed several wee later. If the patient is still hyperthyroid, additional dos are repeated until normal thyroid status is reache These quantities of radioiodine are about 1000 times great as those used for diagnosis of hyperthyroidism, discussed earlier.

HYPOTHYROIDISM

The effects of hypothyroidism in general are oppos to those of hyperthyroidism, but here again, a fe physiological mechanisms peculiar to hypothyroidis alone are involved

Hypothyroidism, like hyperthyroidism, probably al results in most instances from autoimmunity again the thyroid gland, but immunity that destroys the gla rather than stimulating it The thyroid glands of mo of these patients first develop "thyroiditis," which mea thyroid inflammation. This causes progressive deteri ration and finally fibrosis of the gland, with results diminished or absent secretion of thyroid hormone. Ho ever, several other types of hypothyroidism also occu often associated with development of enlarged thyro glands, called thyroid goiter

Endemic Colloid Goiter. The term goiter means a greatly enlarged thyroid gland As pointed out in the discussion of iodine metabolism, about 50 mg of iodine is necessary each year for tha formation of adequate quantities of thyroid hormone. In certain areas of the world, notably in the Swiss Alps, in the Andes, and in the Great Lakes region of tha United States, insufficient iodine is present in the soil for the foodstuffs to contain even this minute quantity of iodine Therefore, in days prior to iodized table salt, many persons living in these areas developed extremely large thyroid glands called *endemic goiters*

The mechanism for development of the large endemic goiters is the following Lack of iodine prevents production of thyroid hormone by the thyroid gland, and, as a result, no hormone is available to inhibit production of TSH by the anterior pituitary; this allows the pituitary to secrete excessively large quantities of TSH. The TSH then causes the thyroid cells to secrete tremendous amounts of thyroglobulin (colloid) into the follicles, and the gland grows larger and larger But unfortunately, because of lack of iodine, increased thyroxine and triiodothyronine production does not occur and therefore does not cause the normal suppression of TSH production by the anterior pituitary. The follicles become tremendous in size, and the thyroid gland may increase to as large as 300 to 500 grams or more.

Idiopathic Nontoxic Colloid Goiter. Enlarged thyroid glands similar to those of endemic colloid goiter frequently also occur in persons who do not have iodine deficiency These goitrous glands may secrete normal quantities of thyroid hormones but more frequently the secretion of hormone is depressed, as in endemic colloid goiter.

The exact cause of the enlarged thyroid gland in patients with idiopathic colloid goiter is not known, but most of these patients show signs of mild thyroiditis; therefore, it has been suggested that the thyroiditis causes slight hypothyroidism, which then leads to in-

creased TSH secretion and progressive growth of the noninflamed portions of the gland. This could explain why these glands usually are very nodular, with some portions of the gland growing while other portions are being destroyed by thyroiditis.

In some persons with colloid goiter, the thyroid gland has an abnormality of the enzyme system required for the formation of the thyroid hormones. Among the abnormalities often encountered are

(1) *Deficient iodide-trapping mechanism, in which* iodine is not pumped adequately into the thyroid cells;

(2) Deficient peroxidase system, in which the iodides are not oxidized to the iodine state,

(3) Deficient coupling of iodinated tyrosines in the thyroglobulin molecule, so that the final thyroid hormones cannot be formed, and

(4) Deficiency of the deiodinase enzyme; this prevents recovery of iodine from the iodinated tyrosines that are not coupled to form the thyroid hormones (this is actually most of the iodine), thus leading to iodine deficiency.

Finally, some foods contain *goitrogenic substances* that have a propylthiouracil-type of antithyroid activity, thus also leading to TSH-stimulated enlargement of the thyroid gland. Such goitrogenic substances are found in some varieties of turnips and cabbages

Characteristics of Hypothyroidism. Whether hypothyroidism is due to thyroiditis, endemic colloid goiter, idiopathic colloid goiter, destruction of the thyroid gland by irradiation, or surgical removal of the thyroid gland, the physiological effects are the same These include fatigue and extreme somnolence with sleeping 14 to 16 hours a day, extreme muscular sluggishness, slowed heart rate, decreased cardiac output, decreased blood volume, sometimes increased weight, constipation, mental sluggishness, failure of many trophic functions in the body evidenced by depressed growth of hair and scaliness of the skin, development of a froglike husky voice, and, in severe cases, development of an edematous appearance throughout the body called myxedema

Myxedema. The patient with almost total lack of thyroid function develops *myxedema. Figure 76–8* shows such a patient, illustrating bagginess under the eyes and swelling of the face. In this condition, for reasons not yet explained, greatly increased quantities of proteoglycans (mucopolysaccharides) containing mainly hyaluronic acid collect in the interstitial spaces, and this causes the total quantity of interstitial fluid also to increase. The fluid is adsorbed to the proteoglycans, thus greatly increasing the quantity of "ground substance" gel in the tissues. Because of the gel nature of the excess fluid, it is relatively immobile, and the edema is nonpitting in type.

Arteriosclerosis in Hypothyroidism. As pointed out in Chapter 68, lack of thyroid hormone increases the quantity of blood lipids, most importantly cholesterol, and the increase in blood cholesterol is usually associated with increased atherosclerosis and arteriosclerosis. Therefore, many hypothyroid patients, particularly those with myxedema, develop arteriosclerosis, which results in peripheral vascular disease, deafness, and often extreme coronary sclerosis with consequent early death.

Diagnostic Tests in Hypothyroidism. The tests already described for diagnosis of hyperthyroidism give the opposite results in hypothyroidism. The free thyroxine in the blood is low. The basal metabolic rate in

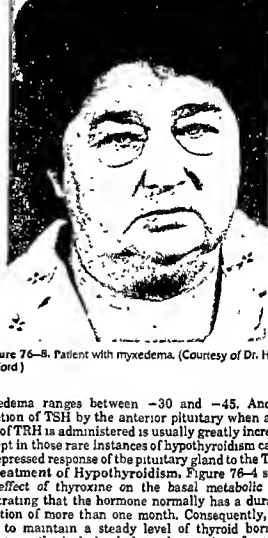


Figure 76–8. Patient with myxedema. (Courtesy of Dr. Herbert Langford.)

myxedema ranges between −30 and −45. And the secretion of TSH by the anterior pituitary when a test dose of TRH is administered is usually greatly increased (except in those rare instances of hypothyroidism caused by depressed response of the pituitary gland to the TRH).

Treatment of Hypothyroidism. Figure 76–4 shows the *effect* of thyroxine on the basal metabolic rate, illustrating that the hormone normally has a duration of action of more than one month. Consequently, it is easy to maintain a steady level of thyroid hormone activity in the body by daily oral ingestion of a tablet or so of thyroid extract or pure thyroxine. Furthermore, proper treatment of the hypothyroid patient results in such complete normality that formerly myxedematous patients properly treated have lived into their 90s after treatment for over 50 years.

Cretinism

Cretinism is the condition caused by extreme hypothyroidism during fetal life, infancy, and childhood, and it is characterized especially by failure of growth Cretinism results from congenital lack of a thyroid gland (*congenital cretinism*), from failure of the thyroid gland to produce thyroid hormone because of a genetic deficiency of the gland, or from iodine lack in the diet (*endemic cretinism*). The severity of endemic cretinism varies greatly, depending on the amount of iodine in the diet, and whole populaces of an endemic area have been known to have cretinoid tendencies.

A newborn baby without a thyroid gland may have absolutely normal appearance and function because be had been supplied with some (but usually not enough) thyroid hormone by the mother while *in utero*, but a few weeks after birth his movements become sluggish, and both his physical and mental growth are greatly retarded. Treatment of the cretin at any time usually

causes normal return of physical growth, but, unless the cretin is treated within a few months after birth, his mental growth will be permanently retarded, and even then this might be too late for optimal therapy. This is probably due to the fact that physical development of the neuronal cells of the central nervous system is rapid during the latter part of fetal life and also the first year of life so that any retardation at this point is extremely detrimental.

Skeletal growth in the cretin is characteristically more inhibited than is soft tissue growth. As a result of this disproportionate rate of growth, the soft tissues are likely to enlarge excessively, giving the cretin the appearance of an obese and stocky, short child. Indeed, occasionally the tongue becomes so large in relation to the skeletal growth that it obstructs swallowing and breathing, inducing a characteristic guttural breathing that sometimes chokes the baby.

REFERENCES

Bayliss, R. I. Thyroid Disease. New York, Oxford University Press, 1982.

Burrow, G. N. The thyroid gland and reproduction. *In* Yen, S. S. C., and Jaffe, R. B. (eds.) Reproductive Endocrinology. Philadelphia, W. B. Saunders Co., 1978, p. 373.

Chopra, I. J., and Solomon, D. H. Pathogenesis of hyperthyroidism. *Annu. Rev. Med.*, *34*:267, 1983.

DeGroot, L. J. Thyroid hormone action. *In* DeGroot, L. J., *et al.* (eds.) Endocrinology, Vol. 1. New York, Grune & Stratton, 1979, p. 357.

DeGroot, L. J., and Taurog, A. Secretions of thyroid hormone. *In* DeGroot, L. J., *et al.* (eds.) Endocrinology, Vol. 1. New York, Grune & Stratton, 1979, p. 343.

Dumont, J. E., and Vassart, G. Thyroid gland metabolism and the action of TSH. *In* DeGroot, L. J., *et al.* (eds.) Endocrinology, Vol. 1. New York, Grune & Stratton, 1979, p. 311.

Edelman, I. S. Thyroid thermogenesis. *N. Engl. J. Med.*, *290*:1303, 1974.

Ekins, R., *et al.* (eds.) Free Thyroid Hormones. New York, Excerpta Medica, 1979.

Ermans, A. M. Endemic goiter and endemic cretinism. *In* DeGroot, L. J., *et al.* (eds.) Endocrinology, Vol. 1. New York, Grune & Stratton, 1979, p. 501.

Greenberg, A. H., *et al.* Effects of thyroid hormone on growth, differentiation and development. *In* Greep, R. O., and Astwood, E. B. (eds.) Handbook of Physiology, Sec. 7, Vol. 3. Baltimore, Williams & Wilkins, 1974, p. 377.

Greer, M. A., and Hailbach, H. Thyroid secretion. *In* Greep, R. O., and Astwood, E. B. (eds.) Handbook of Physiology, Sec. 7, Vol. 3. Baltimore, Williams & Wilkins, 1974, p. 135.

Hennessy, J. F. Hypothyroidism. Discussions in Patient Management. Garden City, N.Y., Medical Examination Publishing Co., 1978.

Ho, K. L. Basophilic degeneration of skeletal muscle in hypothyroid myopathy. Histochemical and ultrastructural studies. *Arch. Pathol. Lab. Med.*, *108*:239, 1984.

Jensen, E. V., and DeSombre, E. R. Steroid hormone receptors and action. *In* DeGroot, L. J., *et al.* (eds.) Endocrinology, Vol. 3. New York, Grune & Stratton, 1979, p. 2055.

King, D. B., and May, J. D. Thyroidal influence on body growth. *J. Exp. Zool.*, *232*:453, 1944.

Kourides, I. A., *et al.* The regulation and organization of thyroid stimulating hormone genes. *Recent Prog. Horm. Res.*, *40*:79, 1984.

Krieger, D. T. (ed.) Current Therapy in Endocrinology, 1983–1984. St. Louis, C. V. Mosby, 1983.

Lenzen, S., and Bailey, C. J. Thyroid hormones, gonadal and adrenocortical steroids and the function of the islets of Langerhans. *Endocr. Rev.*, *5*:411, 1984.

Li, C. H. (ed.) Thyroid Hormones. New York, Academic Press, 1978.

Martin, J. B. Regulation of the pituitary thyroid axis. *In* MTP International Review of Science, Physiology, Vol. 5. Baltimore, University Park Press, 1974, p. 67.

Matovinovic, J. Endemic goiter and cretinism at the dawn of the third millennium. *Annu. Rev. Nutr.*, *3*:341, 1983.

McClung, M. R., and Greer, M. A. Treatment of hyperthyroidism. *Annu. Rev. Med.*, *31*:385, 1980.

McKenzie, J. M., and Zakarya, M. Hyperthyroidism. *In* DeGroot, L. J., *et al.* (eds.) Endocrinology, Vol. 1. New York, Grune & Stratton, 1979, p. 429.

Middleswarth, L. V. Metabolism and excretion of thyroid hormones. *In* Greep, R. O., and Astwood, E. B. (eds.) Handbook of Physiology, Sec. 7, Vol. 3. Baltimore, Williams & Wilkins, 1974, p. 215.

Oppenheimer, J. H. Thyroid hormone action at the nuclear level. *Ann. Intern. Med.*, *102*:374, 1985.

Ramsden, D. B. Peripheral Metabolism and Action of Thyroid Hormones. Montreal, Eden Press, 1977.

Refetoff, S. Thyroid function tests. *In* DeGroot, L. J., *et al.* (eds.) Endocrinology, Vol. 1. New York, Grune & Stratton, 1979, p. 387.

Robbins, J., *et al.* The thyroid and iodine metabolism. *In* Bondy, P. K., and Rosenberg, L. E. (eds.) Metabolic Control and Disease, 8th Ed. Philadelphia, W. B. Saunders Co., 1980, p. 1325.

Roth, J. Receptors for peptide hormones. *In* DeGroot, L. J., *et al.* (eds.) Endocrinology, Vol. 3. New York, Grune & Stratton, 1979, p. 2037.

Stanbury, J. B. (ed.) Endemic Goiter and Endemic Cretinism. New York, John Wiley & Sons, 1980.

Stanbury, J. B., and Kroc, R. L. (eds.) Human Development and the Thyroid Gland. New York, Plenum Press, 1973.

Sterling, K., and Lazarus, J. H. The thyroid and its control. *Annu. Rev. Physiol.*, *39*:349, 1977.

Utiger, R. D. Hypothyroidism. *In* DeGroot, L. J., *et al.* (eds.) Endocrinology, Vol. 1. New York, Grune & Stratton, 1979, p. 471.

Wilber, J. F. Human pituitary thyrotropin. *In* DeGroot, L. J., *et al.* (eds.) Endocrinology, Vol. 1. New York, Grune & Stratton, 1979, p. 141.

77

The Adrenocortical Hormones

The *adrenal glands*, each of which weighs about 4 grams, lie at the superior poles of the two kidneys. And, as illustrated in Figure 77–1, each gland is composed of two distinct parts, the *adrenal medulla* and the *adrenal cortex*. The adrenal medulla, the central 20 per cent of the gland, is functionally related to the sympathetic nervous system; it secretes the hormones *epinephrine* and *norepinephrine* in response to sympathetic stimulation. In turn, these hormones cause almost the same effects as direct stimulation of the sympathetic nerves in all parts of the body. These hormones and their effects were discussed in detail in Chapter 57 in relation to the sympathetic nervous system.

The adrenal cortex secretes an entirely different group of hormones, called *corticosteroids*. These hormones are all synthesized from the steroid cholesterol, and they all have similar chemical formulas However, very slight differences in their molecular structures give them several very different but very important functions.

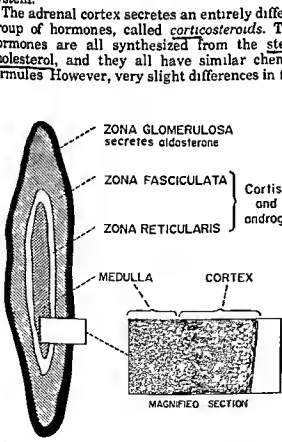


Figure 77–1. Secretion of adrenocortical hormones by the different zones of the adrenal cortex.

Mineralocorticoids and **Glucocorticoids.** Two major types of adrenocortical hormones, the *mineralocorticoids* and the *glucocorticoids*, are secreted by the adrenal cortex. In addition to these, small amounts of sex hormones are secreted, especially *androgenic hormones*, which exhibit approximately the same effects in the body as the male sex hormone testosterone. These are normally of only slight importance, though in certain abnormalities of the adrenal cortices extreme quantities can be secreted (which is discussed later in the chapter) and can then result in masculinizing effects

The *mineralocorticoids* have gained this name because they especially affect the electrolytes of the extracellular fluids—sodium and potassium, in particular The *glucocorticoids* have gained their name because they exhibit an important effect in increasing blood glucose concentration. However, the glucocorticoids have additional effects on both protein and fat metabolism which likely are equally as important if not more important to body function than are their effects on carbohydrate metabolism.

Over 30 different steroids have been isolated from the adrenal cortex, but only two of these are of exceptional importance to the endocrine function of the human body—*aldosterone*, which is the principal mineralocorticoid, and *cortisol*, which is the principal glucocorticoid.

CHEMISTRY OF ADRENOCORTICAL SECRETION

The Layers of the Adrenal Cortex, and Hormone Formation. Figure 77–1 shows that the adrenal cortex is composed of three relatively distinct layers. *Aldosterone* is secreted by the *zona glomerulosa*, the outermost and very thin layer on the surface. *Cortisol* and several other glucocorticoids are secreted by both the *zona fasciculata*, the middle layer, and the *zona reticularis*, the deep layer, with probably more secretion of these hormones by the zona fasciculata than reticularis. The

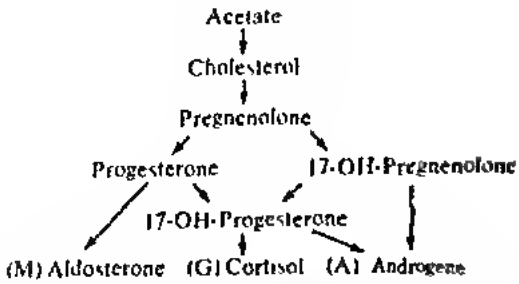


Figure 77–2. The major steps in the synthesis of the three principal adrenal steroids. The physiologic characteristics are designated (M) mineralocorticoid effect, (G) glucocorticoid effect and (A) androgenic effect.

adrenal androgens are also secreted by both these layers, but probably more by the zona reticularis.

Those conditions that increase the output of aldosterone cause hypertrophy of the zona glomerulosa while not affecting the other two zones. On the other hand, those factors that cause increased secretion of cortisol and adrenal androgens cause hypertrophy of the zona fasciculata and zona reticularis while having very little or no effect on the zona glomerulosa; this is especially true of stimulation of the gland by ACTH, as will be discussed later in this chapter.

Chemistry of the Adrenocortical Hormones. All the adrenocortical hormones are steroid compounds. They are formed mainly from cholesterol absorbed directly from the circulating blood by endocytosis through the cell membrane. This membrane has specific receptors for the low density lipoproteins that contain very high concentrations of cholesterol, and attachment of these lipoproteins to the membrane promotes the endocytotic process. Small amounts of cholesterol are also synthesized within the cortical cells from acetyl coenzyme A, this too can be used for forming the adrenocortical hormones.

Figure 77–2 gives the principal steps in the formation of the three important steroid products of the adrenal cortex: aldosterone, cortisol, and the androgens. Essentially all these steps occur in two of the organelles of the cell, the *mitochondria* and the *endoplasmic reticulum*, some steps occurring in one of these organelles and some in the other. Each step is catalyzed by a specific enzyme system. A change in even a single enzyme in the schema can cause vastly different types of hormones to be formed, such as especially large quantities of masculinizing or, very rarely, feminizing sex hormones or other steroid compounds not normally present in the blood but that have either mineralocorticoid or glucocorticoid actions or a combination of both.

Figure 77–3 illustrates the chemical formulas of aldosterone and cortisol. It is the oxygen atom bound at the number 18 carbon of the cholesterol nucleus that is most important in providing the mineralocorticoid activity of aldosterone. The glucocorticoid activity of cortisol is provided principally by the presence of the keto-oxygen on carbon number 3 and the hydroxylation of carbon numbers 11 and 21.

In addition to aldosterone and cortisol, which respectively are the principal mineralocorticoid and glucocorticoid hormones, still other steroids having one or both of these activities are secreted in small amounts even normally by the adrenal cortex. And several additional potent steroid hormones not normally formed in the adrenal glands have been synthesized and are used in various forms of therapy. The more important of the adrenocortical hormones are the following:

Mineralocorticoids:

Aldosterone (very potent, accounts for 95 per cent or more of mineralocorticoid activity)
Desoxycorticosterone (one fifteenth as potent as aldosterone, very small quantities secreted)
Corticosterone (slight activity)
9α-Fluorocortisol (synthetic, slightly more potent than aldosterone)
Cortisol (very slight activity)
Cortisone (synthetic, slight activity)

Glucocorticoids

Cortisol (very potent, accounts for about 95 per cent of all glucocorticoid activity)
Corticosterone (about 4 per cent of total glucocorticoid activity, but much less potent than cortisol)
Cortisone (synthetic, almost as potent as cortisol)
Prednisone (synthetic, four times as potent as cortisol)
Methylprednisone (synthetic, five times as potent as cortisol)
Dexamethasone (synthetic, 30 times as potent as cortisol)

It is clear from this list that some of these hormones have both glucocorticoid and mineralocorticoid activities. It is especially significant that cortisol has a small amount of mineralocorticoid activity because some syndromes of excess cortisol secretion can cause significant mineralocorticoid effects.

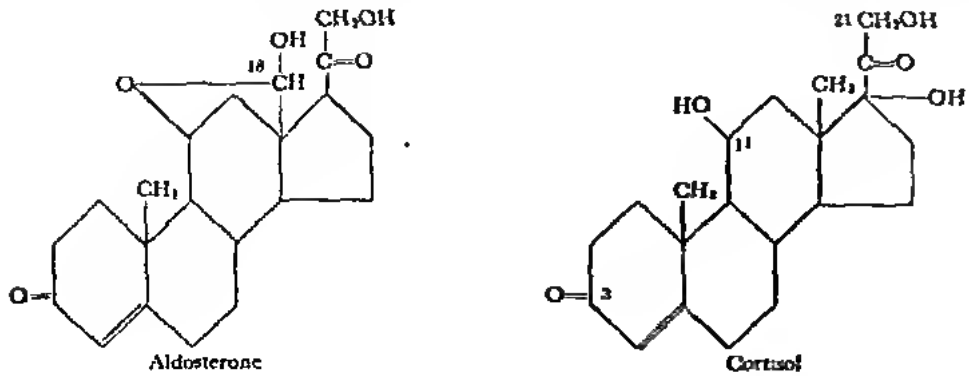


Figure 77–3. The two important corticosteroids.

The very intense glucocorticoid activity of dexamethasone, which has almost zero mineralocorticoid activity, makes this an especially important drug for stimulating specific glucocorticoid activity.

Transport and Fate of the Adrenal Hormones. Cortisol combines with a globulin called *cortisol-binding globulin* or *transcortin* and, to a lesser extent, with albumin—about 94 per cent is normally transported in the bound form and about 6 per cent free. On the other hand, aldosterone combines only loosely with the plasma proteins so that about 50 per cent is in the free form. In both the combined and free forms the hormones are transported throughout the extracellular fluid compartment. In general, the hormones become fixed in the target tissues within an hour or two for cortisol and within about 30 minutes for aldosterone.

The adrenal steroids are degraded mainly in the liver and conjugated especially to form glucuronides and, to a lesser extent, sulfates. About 25 per cent of these are excreted in the bile and then in the feces and the remaining 75 per cent in the urine. The conjugated forms of these hormones are inactive.

The normal concentration of aldosterone in blood is about 6 nanograms (6 billionths of a gram) per dl, and the secretory rate is about 150 μg per day.

The concentration of cortisol in the blood averages 12 μg/dl, and the secretory rate averages 15 mg per day.

FUNCTIONS OF THE MINERALOCORTICOIDS—ALDOSTERONE

Total loss of adrenocortical secretion usually causes death within three days to two weeks unless the person receives extensive salt therapy or mineralocorticoid therapy. Without mineralocorticoids, the potassium ion concentration of the extracellular fluid rises markedly, the sodium and chloride concentrations decrease, and the total extracellular fluid volume and blood volume also become greatly reduced. The person soon develops diminished cardiac output, which proceeds to a shocklike state followed by death. This entire sequence can be prevented by the administration of aldosterone or some other mineralocorticoid. Therefore, the mineralocorticoids are said to be the acute "life-saving" portion of the adrenocortical hormones, while the glucocorticoids are equally necessary to allow the person to resist the destructive effects of life's intermittent "stresses," as is discussed later in the chapter.

Aldosterone exerts at least 95 per cent of the mineralocorticoid activity of the adrenocortical secretion, but cortisol, the major glucocorticoid secreted by the adrenal cortex, also provides a small amount of mineralocorticoid activity. Other adrenal steroids secreted in small amounts that have mineralocorticoid effects are *corticosterone*, which also exerts glucocorticoid effects, and *deoxycorticosterane*, which has almost the same effects as aldosterone but with a potency one fifteenth that of aldosterone.

RENAL EFFECTS OF ALDOSTERONE

By far the most important function of aldosterone is to promote transport of sodium and potassium through the renal tubular walls and, to a lesser extent, transport of hydrogen ions. The mechanisms of these effects were discussed in detail in Chapters 34 through 37. However, let us summarize briefly the renal and body fluid effects of aldosterone.

Effect on Tubular Reabsorption of Sodium and Tubular Secretion of Potassium. It will be recalled from Chapter 34 that aldosterone causes an exchange transport of sodium and potassium—that is, absorption of sodium and simultaneous excretion of potassium by the tubular epithelial cells—in the distal tubule, collecting tubule, and collecting duct. Therefore, aldosterone causes sodium to be conserved in the extracellular fluid while potassium is excreted into the urine.

A high concentration of aldosterone in the plasma can decrease the sodium loss into the urine to as little as a few milligrams a day. At the same time, potassium loss into the urine increases manyfold. Conversely, total lack of aldosterone secretion can cause loss of as much as 20 grams of sodium in the urine a day, an amount equal to one fifth of all the sodium in the body. But, at the same time, potassium is conserved tenaciously in the extracellular fluid.

Therefore, the net effect of excess aldosterons in the plasma is to increase the total quantity of sodium in the extracellular fluid while decreasing the potassium. Yet, strangely enough, a large increase in aldosterone increases the sodium ion *concentration* in the extracellular fluid very little. Instead, it increases the extracellular fluid volume considerably. These differences can be explained as follows:

Mild Effect of Aldosterone on Sodium Ion Concentration. When aldosterone causes excess sodium reabsorption, this also causes reabsorption of almost an equivalent amount of water, an effect that was described in Chapter 36. The main reason is that, when sodium is absorbed through the tubular epithelium, an osmotic gradient is created from the tubules toward the peritubular fluids, and water then "follows" the sodium. Secondly, if not enough water follows, the sodium concentration in the extracellular fluid rises slightly and elicits increased antidiuretic hormone secretion as well as increased thirst; these two together then enhance the amount of water in the extracellular fluid. Consequently, even vast amounts of aldosterone will rarely raise the sodium ion concentration more than 2 to 3 per cent, and total lack of aldosterone causes only a 5 to 8 per cent decrease in sodium concentration.

Effect on Extracellular Fluid Volume. Even though aldosterone increases the sodium ion concentration very little, the combined, almost equal absorption of sodium and water increases

cellular fluid volume much more, increasing the volume to as high as 10 to 20 per cent above normal with great excesses of aldosterone; or the extracellular fluid may decrease to as low as 20 to 25 per cent below normal in the absence of aldosterone. Consequently, the extracellular fluid volume tends to change in proportion to the rate of aldosterone secretion, but only within a narrow range.

Aldosterone Escape. When aldosterone is first administered, sodium retention and increase in extracellular fluid volume become maximal within about 3 days. Then during the next few days, diuresis and natriuresis occur, and most of the retained sodium and excess extracellular fluid volume is lost in the urine, leaving a final increase in extracellular fluid volume rarely exceeding 10 per cent. This secondary loss of sodium and extracellular fluid is called *aldosterone escape*. The precise mechanism of this escape is not known. Experiments in our laboratory have suggested that it results from a rise in arterial blood pressure of 10 to 15 mm Hg resulting from the increased fluid volume. This pressure rise is enough to cause the pressure diuresis and natriuresis phenomena, as was explained in Chapters 22 and 36. However, other physiologists have suggested that some unknown hormonal changes are responsible, though the exact hormones involved have not been discovered.

Hypokalemia and Muscle Weakness; Hyperkalemia and Cardiac Toxicity. The excessive loss of potassium ions from the extracellular fluid into the urine under the influence of aldosterone causes serious decrease in the plasma potassium concentration, often decreasing it from the normal value of 4.5 mEq/liter to as low as 1 to 2 mEq/liter. This condition is called *hypokalemia*. When the potassium ion concentration falls below approximately one half normal, severe muscle weakness often develops. This is caused by alteration of the electrical properties of the nerve and muscle fiber membranes (see Chapter 10), which prevents transmission of action potentials.

On the other hand, when aldosterone is deficient, the extracellular fluid potassium ion concentration can rise far above normal. When it rises to approximately double normal, serious cardiac toxicity, including weakness of heart contraction and development of arrhythmia, becomes evident, a slightly higher concentration of potassium leads inevitably to a cardiac death.

Effect of Aldosterone on Increasing Tubular Hydrogen Ion Secretion, with Resultant Mild Alkalosis. Though aldosterone mainly causes potassium to be secreted into the tubules in exchange for sodium reabsorption, to a much smaller extent it also causes tubular secretion of hydrogen ions in exchange for sodium. The obvious effect of this is to decrease the hydrogen ion concentration in the extracellular fluid. However, this effect is not a strong one, usually causing only a mild degree of alkalosis.

Effect of Aldosterone on Circulatory Function. The circulatory effects of aldosterone result almost entirely from the increase in extracellular fluid volume. In the absence of aldosterone secretion, with a decrease in extracellular fluid volume to 20 to 25 per cent below normal and a comparable decrease in plasma volume, circulatory shock develops rapidly. Indeed, in complete lack of aldosterone, a person not treated with extra intake of salt and/or administration of a mineralocorticoid drug is likely to die of circulatory shock within as few as four to eight days.

In the case of hypersecretion of aldosterone, not only is the extracellular fluid volume increased but the blood volume and cardiac output as well. Each of these can increase to as much as 20 to 30 per cent above normal in the first few days of excess aldosterone secretion, but after aldosterone escape occurs, the volumes and cardiac output usually return to no more than 5 to 10 per cent above normal. Nevertheless, over a prolonged period of time even these small increases are sufficient to cause moderate to severe hypertension, as we shall discuss later in the chapter in relation to primary aldosteronism.

EFFECTS OF ALDOSTERONE ON SWEAT GLANDS, SALIVARY GLANDS, AND INTESTINAL ABSORPTION

Aldosterone has almost the same effects on sweat glands and salivary glands that it has on the renal tubules. Both these glands form a primary secretion that contains large quantities of sodium chloride, but much of the sodium chloride on passing through the excretory ducts is reabsorbed whereas potassium and bicarbonate ions are secreted. Aldosterone greatly increases the reabsorption of sodium chloride and the secretion of potassium by the ducts. The effect on the sweat glands is important to conserve body salt in hot environments, and the effect on the salivary glands is necessary to conserve salt when excessive quantities of saliva are lost.

Aldosterone also greatly enhances sodium absorption by the intestines, which obviously prevents loss of sodium in the stools. On the other hand, in the absence of aldosterone, sodium absorption from the intestine can be very poor, leading to failure to absorb anions and water as well. The unabsorbed sodium chloride and water then lead to diarrhea, with further loss of salt from the body.

CELLULAR MECHANISM OF ALDOSTERONE ACTION

Although for many years we have known the overall effects of mineralocorticoids on the body, the basic action of aldosterone on the tubular cells to increase transport of sodium is still only partly understood. The sequence of events that leads to increased sodium reabsorption seems to be the following:

First, because of its lipid solubility in the cellular membranes aldosterone diffuses readily to the interior of the tubular epithelial cells.

Second, in the cytoplasm of the tubular cells aldosterone combines with a highly specific cytoplasmic *receptor protein*, a protein that has a ster-

eomolecular configuration that will allow only aldosterone or extremely similar compounds to combine.

Third, the aldosterone-receptor complex diffuses into the nucleus where it may undergo further alterations, and then it induces specific portions of the DNA to form a type or types of messenger RNA related to the process of sodium and potassium transport.

Fourth, the messenger RNA diffuses back into the cytoplasm where, operating in conjunction with the ribosomes, it causes protein formation. The protein formed is one or more enzymes or carrier substances required for sodium transport, possibly a specific ATPase that catalyzes energy transfer from cytoplasmic ATP to the sodium transport mechanism of the cell membrane.

Thus, aldosterone does not have an immediate effect on sodium transport but must await the sequence of events that leads to the formation of the specific intracellular substance or substances required for sodium transport. Approximately 20 to 30 minutes are required before new RNA appears in the cells, and approximately 45 minutes are required before the rate of sodium transport begins to increase; the effect reaches maximum only after several hours.

REGULATION OF ALDOSTERONE SECRETION

The regulation of aldosterone secretion is so deeply intertwined with the regulation of extracellular fluid electrolyte concentrations, extracellular fluid volume, blood volume, arterial pressure, and many special aspects of renal function that it is not possible to discuss the regulation of aldosterone secretion independently of all these other factors. This subject has already been presented in great detail in Chapter 36, to which the reader is referred. However, it is important to list hers as well the most important points of aldosterone secretion control.

Let us note first that aldosterone is secreted by the *zona glomerulosa*, a very thin zone of cells located on the surface of the adrenal cortex immediately beneath the capsule. These cells function almost entirely independently of the deeper cells in the zona reticularis and zona fasciculata, which secrete cortisol and the androgens. Furthermore, the regulation of aldosterone secretion is almost entirely independent of the regulation of these other hormones.

Four different factors are presently known to play essential roles in the regulation of aldosterone. In the probable order of their importance these are:

1. Potassium ion concentration of the extracellular fluid
2. Renin-angiotensin system
3. Quantity of body sodium
4. Adrenocorticotropic hormone (ACTH)

Effect of Potassium Ion Concentration on Aldosterone Secretion. An increase in potassium ion concentration of less than 1 mEq/liter will triple the rate of aldosterone secretion. Furthermore, this secretion will continue at an elevated level indefinitely under the stimulus of excess potassium ions.

This very potent effect of potassium ions is exceedingly important because it establishes a powerful feedback mechanism for control of extracellular fluid potassium ion concentration as follows: (1) An increase in potassium ion concentration causes increased secretion of aldosterone. (2) The aldosterone in turn has a potent effect on the kidneys, causing enhanced excretion of potassium. (3) Therefore, the potassium ion concentration returns either to or nearly to normal. Recent quantitative measurements of the feedback gain of this system show it to be by far the most potent of all the factors controlling aldosterone secretion, as has already been discussed in detail in Chapter 36. Serious interference with this control system can lead to a cardiac death if the potassium concentration rises too high or to a paralytic death if the potassium concentration falls too low.

This effect of potassium ions on aldosterone secretion is a direct effect of the potassium ions on the zona glomerulosa cells, though the cellular mechanism of the effect is unknown.

Effect of the Renin-Angiotensin System on Aldosterone Secretion. Infusion of large amounts of angiotensin into an animal can cause acute increases in aldosterone secretion of as much as eightfold. However, if the angiotensin infusion is continued, the rate of aldosterone secretion falls in about 12 hours to only 50 to 100 per cent above normal. This effect is illustrated in Figure 77–4, which shows the effect of two different angiotensin infusion rates, one of which (the dashed curve) increased the extracellular fluid concentration of angiotensin to about 3 times normal, and the other

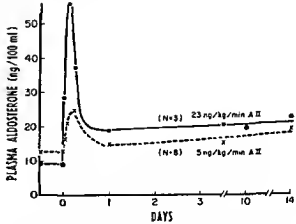


Figure 77–4. Effects on plasma aldosterone concentration caused by continuous infusion of angiotensin II at two different infusion rates for a period of two weeks. Note the very acute effect but the much weaker chronic effect. (Drawn data in Cowley and McCaa, *Circ. Res.*, 39:788, 1976.)

of which (the solid curve) increased the angiotensin concentration to about 15 times normal.

Elevated renin and angiotensin are found very frequently in the same clinical conditions in which aldosterone is also elevated. Therefore, the effect of angiotensin is often stated to be the most important of all the factors controlling aldosterone secretion even though the effect is not as potent as that of increased potassium.

Effect of Decreased Quantity of Total Body Sodium on Aldosterone Secretion. When an animal or human being is placed on a sodium-deficient diet, after several days the rate of aldosterone secretion increases markedly even though the sodium ion concentration of the body fluids does not fall significantly. Suggestions for the cause of this effect have been the following:

1. The diminished sodium leads to diminished extracellular fluid volume, with resultant diminished cardiac output and renal blood flow. This causes enhanced formation of angiotensin, and the angiotensin stimulates aldosterone secretion.
2. Lack of sodium causes retention of potassium by the kidneys, as explained earlier. The elevated potassium could then cause the increased aldosterone secretion.
3. A few experiments have suggested that diminished sodium concentration possibly causes the anterior pituitary gland to secrete some substance (not ACTH) that affects the adrenal glands to increase aldosterone secretion. For the present, this substance is called the *unidentified pituitary factor*.
4. Diminished sodium concentration directly affects the zona glomerulosa cells to enhance aldosterone secretion.

Unfortunately, the importance of each of these mechanisms is yet to be proved. Even so, prolonged sodium depletion is one of the most potent of all the stimulators of aldosterone secretion.

Effect of ACTH on Aldosterone Secretion. ACTH is the anterior pituitary hormone that controls the secretion of glucocorticoids. It also has a *permissive effect* on aldosterone secretion in the following way: All the above regulations of aldosterone secretion will occur if there is only a minimal amount of ACTH present. However, in the total absence of ACTH, the zona glomerulosa of the adrenal cortex partially atrophies and there is almost total atrophy of the remainder of the gland. Thus, total absence of ACTH can lead to a mild to moderate degree of aldosterone deficiency.

FUNCTIONS OF THE GLUCOCORTICOIDS

Even though mineralocorticoids can save the life of an acutely adrenalectomized animal, the animal still is far from normal. Instead, its metabolic systems for utilization of proteins, carbohydrates, and fats are considerably deranged. Furthermore, the animal cannot resist different types of physical or even mental stress, and minor illnesses such as respiratory tract infections can lead to death. Therefore, the glucocorticoids have functions just as important to long-continued life of the animal as do the mineralocorticoids. These are explained in the following sections.

At least 95 per cent of the glucocorticoid activity of the adrenocortical secretions results from the secretion of *cortisol*, known also as *hydrocortisone*. In addition to this, a small amount of glucocorticoid activity is provided by *corticosterone* and a minute amount by *cortisone*.

EFFECTS OF CORTISOL ON CARBOHYDRATE METABOLISM

Stimulation of Gluconeogenesis. By far the best-known metabolic effect of cortisol and other glucocorticoids on metabolism is their ability to stimulate gluconeogenesis by the liver, often increasing the rate of gluconeogenesis as much as six- to tenfold. This results mainly from two different effects of cortisol:

First, all the enzymes required to convert amino acids into glucose are increased in the liver cells. This results from the effect of the glucocorticoids to activate DNA transcription in the liver cell nuclei, in the same way that aldosterone functions in the renal tubular cells, with formation of messenger RNAs that in turn lead to the array of enzymes required for gluconeogenesis.

Second, cortisol causes mobilization of amino acids from the extrahepatic tissues, mainly from muscle. As a result, more amino acids become available in the plasma to enter into the gluconeogenesis process of the liver and thereby to promote the formation of glucose.

One of the effects of increased gluconeogenesis is a marked increase in glycogen in the liver cells.

Decreased Glucose Utilization by the Cells. Cortisol also causes a moderate decrease in the rate of glucose utilization by the cells everywhere in the body. Though the cause of this decrease is unknown, most physiologists believe that somewhere between the point of entry of glucose into the cells and its final degradation cortisol directly delays the rate of glucose utilization. A suggested mechanism is based on the observation that glucocorticoids depress the oxidation of NADH. Since NADH must be oxidized to allow rapid glycolysis, this effect could account for the diminished utilization of glucose by the cells.

Also, it is known that glucocorticoids slightly depress glucose transport into the cells, which could be an additional factor that depresses cellular glucose utilization.

Elevated Blood Glucose Concentration, and Adrenal Diabetes. Both the increased rate of gluconeogenesis and the moderate reduction in

rate of glucose utilization by the cells cause the blood glucose concentration to rise. The increase in concentration is occasionally great enough—50 per cent or more above normal—that the condition is called *adrenal diabetes*, and it has many similarities to pituitary diabetes, which was discussed in Chapter 75. Administration of insulin lowers the blood glucose concentration only a moderate amount in adrenal diabetes, not nearly so much as it does in pancreatic diabetes. On the other hand, insulin causes greater decrease in blood glucose concentration in adrenal diabetes than in pituitary diabetes. Therefore, *pituitary diabetes is said to be weakly insulin sensitive, adrenal diabetes moderately insulin sensitive, and pancreatic diabetes strongly insulin sensitive.*

EFFECTS OF CORTISOL ON PROTEIN METABOLISM

Reduction in Cellular Protein. One of the principal effects of cortisol on the metabolic systems of the body is reduction of the protein stores in essentially all body cells except those of the liver. This is caused both by decreased protein synthesis and increased catabolism of protein already in the cells. Both these effects may possibly result from decreased amino acid transport into extrahepatic tissues, as will be discussed later, but this probably is not the only cause since cortisol also depresses the formation of RNA in many extrahepatic tissues, especially muscle and lymphoid tissue.

In the presence of great excesses of cortisol the muscles can become so weak that the person cannot rise from the squatting position. And the immunity functions of the lymphoid tissue can be decreased to a small fraction of normal.

Increased Liver Protein and Plasma Proteins Caused by Cortisol. Coincidently with the reduced proteins elsewhere in the body, the liver proteins become enhanced. Furthermore, the plasma proteins (which are produced by the liver and then released into the blood) are also increased. Therefore, these are exceptions to the protein depletion that occurs elsewhere in the body. It is believed that this is a possible effect of cortisol in enhancing amino acid transport into liver cells (but not into most other cells) and of enhancement of the liver enzymes required for protein synthesis.

Increased Blood Amino Acids, Diminished Transport of Amino Acids into Extrahepatic Cells, and Enhanced Transport into Hepatic Cells. Recent studies in isolated tissue have demonstrated that cortisol depresses amino acid transport into muscle cells and perhaps into other extrahepatic cells.

Obviously, the decreased transport of amino acids into extrahepatic cells decreases their intracellular amino acid concentrations and as a consequence decreases the synthesis of protein. Yet catabolism of proteins in the cells continues to release amino acids from the already existing proteins, and these diffuse out of the cells to increase the plasma amino acid concentration. Therefore, it is said that *cortisol mobilizes amino acids from the tissues.*

The increased plasma concentration of amino acids, plus the fact that cortisol enhances transport of amino acids into the hepatic cells, could also account for enhanced utilization of amino acids by the liver to cause such effects as (1) increased rate of deamination of amino acids by the liver, (2) increased protein synthesis in the liver, (3) increased formation of plasma proteins by the liver, and (4) increased conversion of amino acids to glucose—that is, enhanced gluconeogenesis.

Thus, it is possible that many of the effects of cortisol on the metabolic systems of the body result mainly from this ability of cortisol to mobilize amino acids, though cortisol also increases the liver enzymes required for the hepatic effects by causing the formation of appropriate messenger RNAs.

EFFECTS OF CORTISOL ON FAT METABOLISM

Mobilization of Fatty Acids. In much the same manner that cortisol promotes amino acid mobilization from muscle, it also promotes mobilization of fatty acids from adipose tissue. This increases the concentration of free fatty acids in the plasma, which also increases their utilization for energy. Cortisol moderately enhances the oxidation of fatty acids in the cells as well; this effect perhaps results secondarily to the reduced availability of glycolytic products for metabolism.

The mechanism by which cortisol promotes fatty acid mobilization is not yet understood. However, part of the effect probably results from diminished transport of glucose into the fat cells. It will be remembered that α-glycerophosphate, which is derived from glucose, is required both for deposition of and maintenance of triglycerides in these cells, and in its absence the fat cells begin to release fatty acids.

The increased mobilization of fats, combined with their increased oxidation in the cells, helps shift the metabolic systems of the cells in times of starvation or other stresses from utilization of glucose for energy to utilization of fatty acids instead. This cortisol mechanism, however, requires several hours to become fully developed—not nearly so rapid or nearly so powerful an effect as a similar shift elicited by a decrease in insulin. Nevertheless, it is probably an important factor for long-term conservation of body glucose and glycogen.

Ketogenic Effects of Cortisol. It is frequently said that cortisol has a *ketogenic effect* because

ketosis usually will not develop unless cortisol is available to cause fat mobilization. However, this ketogenic effect occurs only under certain conditions, such as when insulin also is deficient.

Obesity Caused by Cortisol. Despite the fact that cortisol can cause a moderate degree of fatty acid mobilization from adipose tissue, persons with excess cortisol secretion frequently develop a peculiar type of obesity, with excess deposition of fat in the chest and head regions of the body, giving a buffalo-like torso and a rounded face called a "moon face." Careful experiments have shown that this obesity results from excess stimulation of food intake so that fat is generated in some tissues of the body at a rate even more rapidly than it is mobilized and oxidized.

OTHER EFFECTS OF CORTISOL

Function of Cortisol in Different Types of Stress

It is amazing that almost any type of stress, whether it be physical or neurogenic, will cause an immediate and marked increase in ACTH secretion by the anterior pituitary gland, followed within minutes by greatly increased adrenocortical secretion of cortisol. Some of the different types of stress that increase cortisol release are the following:

1. Trauma of almost any type
2. Infection
3. Intense heat or cold
4. Injection of norepinephrine and other sympathomimetic drugs
5. Surgical operations
6. Injection of necrotizing substances beneath the skin
7. Restraining an animal so that it cannot move
8. Almost any debilitating disease

Thus, a wide variety of nonspecific stimuli can cause marked increase in the rate of cortisol secretion by the adrenal cortex.

Yet, even though we know that cortisol secretion often increases greatly in stressful situations, we still are not sure why this is of significant benefit to the animal. One guess, which is probably as good as any other, is that the glucocorticoids cause rapid mobilization of amino acids and fats from their cellular stores, making these available both for energy and for synthesis of other compounds, including glucose, needed by the different tissues of the body. Indeed, it is well known that when proteins are released from most of the tissue cells, the liver cells can use the mobilized amino acids to form new proteins. It has also been shown in a few instances that damaged tissues which are momentarily depleted of proteins can also utilize the newly available amino acids to form new proteins that are essential to the lives of the cells. Or perhaps the amino acids are also used to synthesize such essential intracellular substances as purines, pyrimidines, and creatine phosphate, which are necessary for maintenance of cellular life and reproduction of new cells.

But all this is mainly supposition. It is supported only by the fact that cortisol usually does not mobilize the basic functional proteins of the cells until almost all other proteins have been released. This preferential effect of cortisol in mobilizing labile proteins could make amino acids available to needy cells to synthesize substances essential to life.

Anti-Inflammatory Effects of Cortisol

When tissues are damaged by trauma, by infection with bacteria, or in almost any other way, they almost always will become inflamed. In some conditions the inflammation is more damaging than the trauma or disease itself. Administration of large amounts of cortisol can usually block this inflammation or even reverse many of its effects once it has begun. Before attempting to explain the way in which cortisol functions to block inflammation, let us first review the basic steps in the inflammation process, which was discussed in more detail in Chapter 5.

Basically, there are five main stages of inflammation: (1) release from the damaged tissue cells of chemical substances that activate the inflammation process—chemicals such as histamine, bradykinin, proteolytic enzymes, and so forth; (2) an increase in blood flow in the inflamed area caused by some of the released products from the tissues, a process called *erythema;* (3) leakage of large quantities of almost pure plasma out of the capillaries into the damaged areas, followed by clotting of the tissue fluid, thus causing a *nonpitting type of edema;* (4) infiltration of the area by leukocytes; and, finally, (5) tissue healing, which is often accomplished at least partially by ingrowth of fibrous tissue.

Yet, unfortunately, we still know very little about the basic mechanisms by which cortisol achieves its *anti-inflammatory effect.* However, very large amounts of cortisol are often required. When such large amounts of cortisol are secreted or are injected into a person, the cortisol has two basic effects: (1) it blocks the early stages of the inflammation process, and (2) if inflammation has already begun, it causes rapid resolution of the inflammation and increased rapidity of healing. These effects are explained as follows.

Prevention of the Development of Inflammation—Lysosome Stabilization and Other Effects. (1) One of the most important anti-inflammation effects of cortisol is its ability to cause *stabilization of the lysosomal membranes.* That is, cortisol makes it much more difficult than is normal for the membranes of the intracellular lysosomes to rupture. Therefore, most of the proteolytic enzymes that are released by damaged cells to cause inflammation, which are mainly stored in

the lysosomes, are released in greatly decreased quantity. Other important effects are:

(2) Cortisol decreases the permeability of the capillaries, which prevents loss of plasma into the tissues and also reduces the migration of white blood cells into the inflamed area.

(3) Cortisol depresses the ability of white blood cells to digest phagocytized tissues, and this blocks further release of inflammatory materials.

(4) Cortisol suppresses the immune system, especially the T cells. Therefore, reduced amounts of T cells and antibodies enter the inflamed area, which reduces tissue reactions that might be caused by these.

(5) Cortisol reduces fever, and this in turn reduces the degree of vasodilatation.

Effect of Cortisol in Causing Resolution of Inflammation. Even after inflammation has become well established, administration of cortisol can often reduce inflammation within hours to several days. The immediate effect is to block most of the factors that are promoting the inflammation. Then, the rate of healing is also enhanced. This probably results from those same, mainly undefined factors that allow the body to resist many other types of physical stress when large quantities of cortisol are secreted: perhaps this results from the mobilization of amino acids and use of these to repair tha damaged tissues; perhaps it results from increased amounts of glucose and fatty acids avallabla for cellular energy; or perhaps it depends on soma catalytic effect of cortisol to inactivate or remova inflammatory products.

Regardless of the precise mechanisms by which the anti-inflammatory effect occurs, this effect of cortisol can play a major role in combatmg certain types of diseases, such as rheumatoid arthritis, rheumatic fever, and acute glomerulonephritis. All these are characterized by severe local inflammation, and the harmful effects on the body are caused mainly by the inflammation itself and not by other aspects of the disease. When cortisol or other glucocorticoids are administered to patients with these diseases, almost invariably the inflammation subsides within 24 to 48 hours. And even though the cortisol does not correct the basic disease condition, merely preventing the damaging effects of the inflammatory response, this alone can be a life-saving measure.

Effect on Allergy. Cortisol blocks the inflammatory response to allergic reactions in exactly the same way that it blocks other types of inflammatory responses. The basic allergic reaction between antigen and antibody is not affected by cortisol, and even some of the secondary effects of the allergic reaction, such as the release of histamine, still occur the same as ever. However, since the inflammatory response is responsible for many of the serious and sometimes lethal effects of allergic reactions, administration of cortisol can be lifesaving. For instance, cortisol effectively prevents shock or death in anaphylaxis, which otherwise kills many persons, as explained in Chapter 6.

Effect on Blood Cells and on Immunity. Cortisol decreases the number of eosinophils and lymphocytes in the blood, this effect begins within a few minutes after the injection of cortisol and is marked within a few hours. Indeed, a finding of lymphocytopenia or eosinopenia is an important diagnostic criterion for overproduction of cortisol by the adrenal gland.

Likewise, administration of large doses of cortisol causes significant atrophy of all the lymphoid tissue throughout the body, which in turn decreases the output of both T cells and antibodies from the lymphoid tissue. As a result, the level of immunity for almost all foreign invaders of the body is decreased. This can occasionally lead to fulminating infection and death from diseases that otherwise would not be lethal, such as fulminating tuberculosis in a person whose disease had previously been arrested. On the other hand, this ability of cortisol and other glucocorticoids to suppress immunity makes these among the most useful of all drugs to prevent immunological rejection of transplanted hearts, kidneys, and other tissues.

Cortisol increases the production of red blood cells, the cause of which is unknown. When excess cortisol is secreted by the adrenal glands, polycythemia often results, and, conversely, when the adrenal glands secrete no cortisol, anemia often results

REGULATION OF CORTISOL SECRETION—ADRENOCORTICOTROPIC HORMONE (ACTH)

Control of Cortisol Secretion by ACTH. Unlike aldosterone secretion by the zona glomerulosa, which is controlled mainly by potassium and angiotensin acting directly on the adrenocortical cells, almost no stimuli have direct effects on the adrenal cells to control cortisol secretion. Instead, secretion of cortisol is controlled almost entirely by *adrenocorticotropic hormone* (ACTH) secreted by the anterior pituitary gland. This hormone, also called *corticotropin* or *adrenocorticotropin*, also enhances the production of adrenal androgens. As pointed out earlier in tho chapter, small amounts of ACTH are also required for aldosterone secretion, providing a permissive role that allows the other, more important factors to exert their more powerful controls.

Chemistry of ACTH. ACTH has been isolated in pure form from the anterior pituitary. It is a large polypeptide having a chain length of 39 amino acids A digested product of ACTH having a chain length of 24 amino acids also has all the trophic effects of the total molecule.

Control of ACTH Secretion by the Hypothalamus—Corticotropin-Releasing Factor (CRF). In the same way that other pituitary hormones are controlled by releasing hormones or factors from the hypothalamus, so also does an important releasing factor control ACTH secretion. This is called *corticotropin-releasing factor* (CRF). It is secreted into the primary capillary plexus of the hypophysial portal system in the median eminence of the hypothalamus and then carried

the anterior pituitary gland where it induces ACTH secretion. The exact chemical nature of CRF has not yet been established, though some experiments suggest that it is a peptide composed of 41 amino acids. The cell bodies of the neurons that secrete CRF are believed to be located in the posterior medial basal hypothalamus. This nucleus in turn receives many nervous connections from the limbic system and lower brain stem. Vasopressin, formed mainly in the supraoptic nuclei and discussed in Chapter 75, can also weakly increase the release of ACTH from the anterior pituitary gland.

The anterior pituitary gland can secrete only small quantities of ACTH in the absence of CRF. Instead, most conditions that cause high ACTH secretory rates initiate this secretion by signals beginning in the hypothalamus and then transmitted by CRF to the anterior pituitary gland.

Mechanism by Which ACTH Activates Adrenocortical Cells to Produce Steroids—Function of Cyclic AMP. The principal effect of ACTH on the adrenocortical cells is to activate *adenyl cyclase* in the cell membrane. This then induces the formation of *cyclic AMP* in the cell cytoplasm, reaching maximum effect in about 3 minutes. The cyclic AMP in turn activates the intracellular enzymes that cause formation of the adrenocortical hormones. Thus, this is another example of cyclic AMP as a *second messenger* hormone.

The most important of all the ACTH-stimulated steps for controlling adrenocortical secretion is activation of the enzyme *desmolase* that causes initial conversion of cholesterol to pregnenolone. This initial conversion is the "rate-limiting" step for all the adrenocortical hormones, which explains why ACTH normally is necessary for any adrenocortical hormones to be formed. Long-term stimulation of the adrenal cortex by ACTH not only increases secretory activity but also causes hypertrophy and proliferation of the adrenocortical cells, especially in the zona fasciculata and reticularis where cortisol and the androgens are secreted.

Effect of Physiological Stress on ACTH Secretion. It was pointed out earlier in the chapter that almost any type of physical or even mental stress can lead within minutes to greatly enhanced secretion of ACTH and consequently of cortisol as well, often increasing cortisol secretion as much as 20-fold. This effect is illustrated forcefully by the curves in Figure 77–5, which shows a manyfold increase in plasma corticosterone concentration in a rat within minutes after the tibia and fibula had been broken (corticosterone is the principal glucocorticoid secreted by the rat adrenal in place of cortisol). It is believed that pain stimuli caused by the stress are first transmitted upward through the brain stem to the periformical area of the hypothalamus and from here into the medial basal hypothalamus and eventually to the median eminence, as shown in Figure 77–6, where CRF is secreted into the hypophysial portal system. Within minutes the entire control sequence leads to large quantities of the glucocorticoids in the blood.

Mental stress can also cause an almost instantaneous increase in ACTH secretion. This is believed to result from increased activity in the limbic system, especially in the region of the amygdala and hippocampus, both of these then transmitting signals to the posterior medial hypothalamus.

Inhibitory Effect of Cortisol on the Hypothalamus and on the Anterior Pituitary Cause Decreased ACTH Secretion. Cortisol has direct negative feedback effects on (1) the hypothalamus to decrease the formation of CRF and (2) the anterior pituitary gland to decrease the formation of ACTH. These feedbacks help regulate the plasma concentration of cortisol. That is, whenever the concentration becomes too great, the feedbacks automatically reduce the ACTH toward normal control level. Or, if the level of cortisol falls too low, lack of the negative feedback will increase the cortisol again toward normal.

Summary of the Control System

Figure 77–6 illustrates the overall system for control of cortisol secretion. The central key to this control is the excitation of the hypothalamus by different types of stress. These activate the entire system to cause rapid release of cortisol, and the cortisol in turn initiates a series of metabolic effects directed toward relieving the damaging nature of the stressful state. In addition, there is also direct feedback of the cortisol to the hypothalamus and anterior pituitary gland to stabilize the concentration of cortisol in the plasma at times

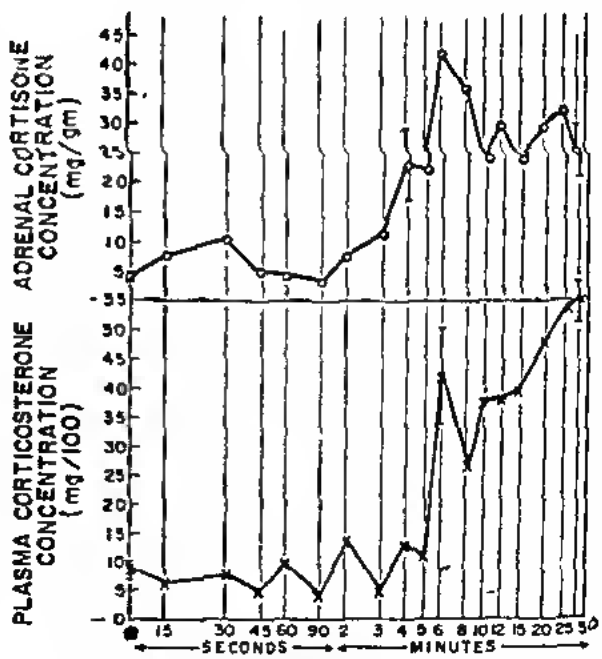


Figure 77–5. Rapid reaction of the adrenal cortex of a rat to stress caused by fracture of the tibia and fibula. (Courtesy of Drs Guillemin, Dear, and Lipscomb.)

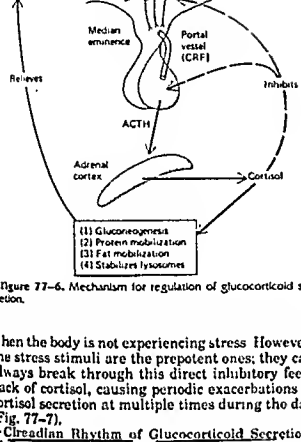


Figure 77–6. Mechanism for regulation of glucocorticoid secretion.

when the body is not experiencing stress However, the stress stimuli are the prepotent ones; they can always break through this direct inhibitory feedback of cortisol, causing periodic exacerbations of cortisol secretion at multiple times during the day (Fig. 77–7).

Circadian Rhythm of Glucocorticoid Secretion. The secretory rates of CRH, ACTH, and cortisol are all high in the early morning but low in the late evening, as also illustrated in Figure 77–7; the plasma cortisol level ranges between a high of about 20 μg/dl an hour before arising in the morning and a low of about 5 μg/dl around midnight. This effect results from a 24-hour cyclic alteration in the signals from the hypothalamus that cause cortisol secretion. When a person changes daily sleeping habits, the cycle changes correspondingly One of the reasons the cycle is so important is that measurements of blood cortisol levels are meaningful only when expressed in terms of the time in the cycle at which the measurements are made.

Secretion of Melanocyte-Stimulating Hormone, Lipotropin, and Endorphin in Association with ACTH

When ACTH is secreted by the anterior pituitary gland, several other hormones that have similar chemical structures are secreted simultaneously. The reason for this is that the RNA molecule that causes ACTH formation initially causes the formation of a considerably larger protein molecule, a preprohormone, that contains ACTH as only one of its subunits. This same preprohormone also contains several other hormones as well, including *alpha-* and *beta-melanocyte-stimulating hormone*, *beta-lipotropin*, and *beta-endorphin*. Under normal conditions, none of these is known to be secreted in enough quantity to have significant effect on the body, but this may not be true when the rate of secretion of ACTH is very high, as occurs in some types of Addison's disease, which will be discussed later.

Beta-endorphin is important because it has an opiate effect on the nervous system; in fact, the substance metenkephalin, which is present in many areas of the brain, is itself a fragment of the beta-endorphin molecule. Some persons believe that the pituitary gland can secrete small amounts of this and other hormones into the hypothalamic tissues as well as into the circulating blood.

The functions of *beta-lipotropin* are yet unknown, but it has been suggested that it stimulates aldosterone secretion in the same way that ACTH stimulates cortisol secretion by the adrenal cortex.

Melanocyte-Stimulating Hormone. Melanocyte-stimulating hormone (MSH) occurs in two forms, an *alpha* and a *beta* form. The alpha form has exactly the same chemical structure as the first 13 amino acids of the 39 amino acid ACTH polypeptide chain.

MSH causes the *melanocytes*, which are located in abundance between the dermis and the epidermis of the skin, to form the pigment *melanin* and to disperse this in the cells of the epidermis. Injection of melanocyte-stimulating hormone into a person over a period of eight to ten days can cause intense darkening of the skin. However, the effect is much greater in persons who have genetically dark skins than in light-skinned persons.

In some lower animals, an intermediate "lobe" of the pituitary gland, called the *pars intermedia*, is highly developed, lying between the anterior and the posterior pituitary lobes. This lobe secretes an especially large amount of melanocyte-stimulating hormone. Furthermore, this secretion is independently controlled by the hypothalamus in response to the amount of light to which the animal is exposed or in response to other environmental factors. For instance, some arctic animals develop darkened fur in the summer and yet have entirely white fur in the winter.

ACTH, because of its similarity to MSH, has about one thirtieth as much melanocyte-stimulating effect as MSH Furthermore, because the quantities of pure MSH secreted in the human being are extremely small whereas those of ACTH are large, it is likely that ACTH is considerably more important than MSH in determining the amount of melanin in the skin

THE ADRENAL ANDROGENS

Several moderately active male sex hormones called *adrenal androgens* (the most important of which is

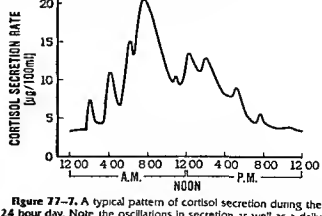


Figure 77–7. A typical pattern of cortisol secretion during the 24 hour day. Note the oscillations in secretion as well as a daily secretory surge an hour or so after awaking in the morning.

dehydroepiandrosterone) are continually secreted by the adrenal cortex, especially so during fetal life as will be discussed more fully in Chapter 82 Also, progestecone and estrogens, which are female sex hormones, are secreted in very minute quantities

In normal physiology of the human being, the adrenal androgens have almost insignificant effects However, it is possible that part of the early development of the male sex organs results from childbood secretion of adrenal androgens. The adrenal androgens also exert mild effects in the female, not only before puberty but also throughout life Much of the growth of the pubic and axillary hair in the female probably results from the action of these hormones Some of the adrenal androgens are converted to testosterone, the major male sex hormone, in the extra-adrenal tissues, which probably accounts for much of their androgenic activity The physiological effects of androgens will be discussed in Chapter 80 in relation to male sexual function

ABNORMALITIES OF ADRENOCORTICAL SECRETION

HYPOADRENALISM—ADDISON'S DISEASE

Addison's disease results from failure of the adrenal cortices to produce adrenocortical hormones, and this in turn is most frequently caused by *primary atrophy* of the adrenal cortices, probably resulting from autoimmunity against the cortices, but also frequently caused by tuberculous destruction of the adrenal glands or invasion of the adrenal cortices by cancer. Basically, the disturbances in Addison's disease are

Mineralocorticoid Deficiency. Lack of aldostecone secretion greatly decreases sodium reabsorption and consequently allows sodium ions, chloride ions, and water to be lost into urine in great profusion. The net result is a greatly decreased extracellular fluid volume Furthermore, the person develops hyperkalemia and mild acidosis because of failure of potassium and hydrogen ions to be secreted in exchange for sodium reabsorption

As the extracellular fluid becomes depleted, the plasma volume falls, the red blood cell concentration rises markedly, the cardiac output decreases, and the patient dies in shock, death usually occurring in the untreated patient four days to two weeks after complete cessation of mineralocorticoid secretion

Glucocorticoid Deficiency. Loss of cortisol secretion makes it impossible for the person with Addison's disease to maintain normal blood glucose concentration between meals because he cannot synthesize significant quantities of glucose by gluconeogenesis Furthermore, lack of cortisol reduces the mobilization of both proteins and fats from the tissues, thereby depressing many other metabolic functions of the body. This sluggishness of energy mobilization when cortisol is not available is one of the major detrimental effects of glucocorticoid lack. However, even when excess quantities of glucose and other nutrients are available, the person's muscles are still weak, indicating that glucocorticoids are also needed to maintain other metabolic functions of the tissues besides simply energy metabolism.

Lack of adequate glucocorticoid secretion also makes the person with Addison's disease highly susceptible to the deteriorating effects of different types of stress, and :n a mild respiratory infection can sometimes cause

Melanin Pigmentation. Another characteristic o most persons with Addison's disease is melanin pigmen tation of the mucous membranes and skin. This melani is not always deposited evenly but occasionally i blotches and especially in the thin skin areas, such a the mucous membranes of the lips and the thin skin c the nipples.

The cause of the melanin deposition is believed to b the following When cortisol secretion is depressed, th normal negative feedback to the hypothalamus an anterior pituitary is also depressed, now allowing tr mendous rates of ACTH secretion as well as simulta neous secretion of increased amounts of melanocyte stimulating hormone (MSH) Probably the tremendou amounts of ACTH cause most of the pigmenting effe because these can stimulate formation of melanin b the melanocytes in the same way that MSH does Eve though MSH has 30 times as much melanocyte-stimu lating effect as ACTH, the amounts secreted by th human being are extremely small.

Diagnosis of Addison's Disease. A person with this disease has almost no urinary secretion of steroids derived from the adrenals Also, immunoassay of the plasma aldosterone and competitive binding assay (see Chapter 74) of cortisol show either very low or unmeasurable values In full-blown Addison's disease, less than one tenth of the normal amount of 17-hydroxysteroids, which are derived almost entirely from cortisol secreted by the adrenals, is excreted in the urine in a 24-hour period However, to be certain that the person has Addison's disease, ACTH is infused into the patient over a period of 8 hours or more, and the plasma cortisol level again measured If the Addison's disease is caused by total destruction of the adrenal cortices, there will be no increase If, on the other hand, the Addison's disease is caused by failure of the anterior pituitary to stimulate the adrenal cortices, administration of ACTH several days in a row will cause the adrenal glands eventually to begin secreting corticosteroids

Treatment of Persons with Addison's Disease. The untreated person with total adrenal destruction dies within a few days to a few weeks because of consuming weakness and eventual circulatory shock. Yet such a person can usually live for years if small quantities of mineralocorticoids and glucocorticoids are administered daily—about 0 1 milligram of fludrocortisone and 30 milligrams of cortisol.

The Addisonian Crisis. As noted earlier in the chapter, great quantities of glucocorticoids are occasionally secreted in response to different types of physical or mental stress In the person with Addison's disease, the output of glucocorticoids does not increase during stress Yet whenever different types of trauma, disease, or other stresses such as surgical operations supervene, a person is likely to develop an acute need for excessive amounts of glucocorticoids, and must be given as much as ten or more times the normal quantities of glucocorticoids in order to prevent death

This critical need for extra glucocorticoids and the associated severe debility in times of stress is called an addisonian crisis

HYPERADRENALISM—CUSHING'S SYNDROME

Hypersecretion of cortisol by the adrenal cortex causes a complex of hormonal effects called Cushing's disease, and this results from either a cortisol-secreting tumor

of one adrenal cortex or general hyperplasia of both adrenal cortices. The hyperplasia in turn is usually caused by increased secretion of ACTH by the anterior pituitary or by "ectopic secretion" of ACTH by a tumor elsewhere in the body, such as an abdominal carcinoma. *Most abnormalities of Cushing's syndrome are ascribable* to abnormal amounts of cortisol, but secretion of androgens is also of significance.

A special characteristic of Cushing's disease is mobilization of fat from the lower part of the body, with concomitant extra deposition of fat in the thoracic and upper abdominal regions, giving rise to a so-called "buffalo" torso. The excess secretion of steroids also leads to an edematous appearance of the face, and the androgenic potency of some of the hormones sometimes causes acne and hirsutism (excess growth of facial hair) The total appearance of the face is frequently described as a "moon face," as illustrated to the left in Figure 77–8, which shows a patient with Cushing's syndrome prior to treatment. About 80 per cent of the patients have hypertension, presumably because of the slight mineralocorticoid effects of cortisol (because a pure glucocorticoid such as dexamethasone actually causes decreased arterial pressure).

Effects on Carbohydrate and Protein Metabolism. The abundance of cortisol secreted in Cushing's syndrome can cause increased blood glucose concentration, sometimes to values as high as 200 mg per cent after meals. This effect results mainly from enhanced gluconeogenesis. If this "adrenal diabetes" lasts for many months, the beta cells in the islets of Langerhans occasionally "burn out" because the high blood glucose greatly overstimulates them to secrete insulin. The destruction of these cells then causes frank diabetes mellitus, which is permanent for the remainder of life.

The effects of glucocorticoids on protein catabolism are often profound in Cushing's syndrome, causing greatly decreased tissue proteins almost everywhere in the body with the exceptions of the liver and the plasma proteins. The loss of protein from the muscles in particular causes severe weakness. The loss of protein synthesis in the lymphoid tissues leads to a suppressed immunity system, so that many of these patients die of infections. Even the collagen fibers in the subcutaneous tissue are diminished so that the subcutaneous tissues *tear easily, resulting in development of large purplish striae* where the subcutaneous tissues have torn apart. In addition, lack of protein deposition in the bones causes osteoporosis with consequent weakness of the bones.

Diagnosis and Treatment of Cushing's Disease. Diagnosis of Cushing's disease is most frequently made on the basis of such typical findings as the buffalo torso, puffiness of the face, mild masculinizing effects, elevated blood glucose concentration that is moderately insulin resistant, several times normal plasma levels of cortisol, and secretion of three to five times the normal quantities of 17-hydroxysteroids in the urine

Treatment in Cushing's disease consists of removing an adrenal tumor if this is the cause, or of decreasing the secretion of ACTH, if this is possible Hypertrophied pituitary glands or even small tumors in the pituitary that oversecrete ACTH can be surgically or microsurgically removed or destroyed by radiation If ACTH secretion cannot easily be decreased, the only satisfactory treatment is usually bilateral total or partial adrenalectomy, followed by administration of adrenal steroids to make up for any insufficiency that develops

PRIMARY ALDOSTERONISM

Occasionally a small tumor of the zona glomerulosa cells occurs and secretes large amounts of aldosterone In a few instances, hyperplastic adrenal cortices secrete aldosterone rather than cortisol The effects of the excess aldosterone are those discussed in detail earlier in the chapter The most important effects are hypokalemia, slight increase in extracellular fluid volume and blood volume, very slight increase in plasma sodium concentration (usually not over a 2 to 3 per cent increase), and almost always hypertension Especially interesting in

Figure 77–8. A person with Cushing's disease before subtotal adrenalectomy (left) and at adrenalectomy (right) (Courtesy of Dr Leonard Posey)

primary aldosteronism are occasional periods of muscular paralysis caused by the hypokalemia The paralysis is caused by a depressant effect of hypokalemia on action potential transmission by the nerve fibers, as was explained in Chapter 10.

One of the diagnostic criteria of primary aldosteronism is a decreased plasma renin concentration This results from feedback suppression of renin secretion caused by the excess aldosterone or by the excess extracellular fluid volume and arterial pressure resulting from the aldosteronism

Treatment of primary aldosteronism is usually surgical removal of the tumor or of most of the adrenal tissue when hyperplasia is the cause In about 30 per cent of the treated patients, the arterial pressure fails to return to normal even though plasma aldosterone concentrations do become normal; this presumably results from some long-term damage to the kidneys.

ADRENOGENITAL SYNDROME

An occasional adrenocortical tumor secretes excessive quantities of androgens that cause intense masculinizing effects throughout the body. If this occurs in a female, she develops virile characteristics, including growth of a beard, a much deeper voice, occasionally baldness if she also has the genetic inheritance for baldness, masculine distribution of hair on the body and on the pubis, growth of the clitoris to resemble a per and deposition of proteins in the skin and especially the muscles to give typical masculine characteristics

In the prepubertal male a virilizing adrenal tum causes the same characteristics as in the female, p rapid development of the male sexual organs and cr tion of male sexual desires. Typical development of t male sexual organs in a 4 year old boy with the adr ogenital syndrome is shown in Figure 77–9.

In the adult male, the virilizing characteristics of the adrenogenital syndrome are usually completely obscured by the normal virilizing characteristics of the testosterone secreted by the testes. Therefore, it is often difficult to make a diagnosis of adrenogenital syndrome in the male adult.

In the adrenogenital syndrome, the excretion of 17-ketosteroids (which are derived from androgens) in the urine may be as much as 10 to 15 times normal

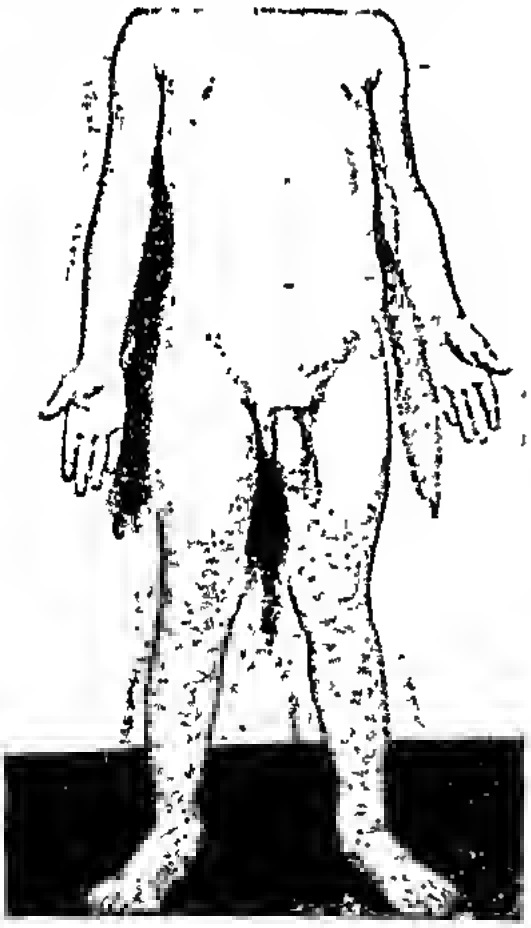


Figure 77–9. Adrenogenital syndrome in a 4 year old boy (Courtesy of Dr. Leonard Posey)

REFERENCES

Bondy, P K The adrenal cortex *In* Bondy, P K, and Rosenberg, L E (eds) Metabolic Control and Disease, 8th Ed Philadelphia, W B Saunders Co 1980, p 1427

Chrousos, G P, *et al* NIH Conference Clinical applications of corticotropin releasing factor *Ann Intern Med*, *102* 344, 1985

Fanestil, D D, and Park, C S Steroid hormones and the kidney *Annu Rev Physiol*, *43* 637, 1981

Fevold, H R Regulation of the adrenal and gonadal microsomal mixed function oxygenases of steroid hormone biosynthesis *Annu Rev Physiol* *45* 19, 1983

Gann, D S *et al* Neural interaction in control of adrenocorticotropin *Fed Proc* *44* 161 1985

Groner B, *et al* DNA sequences involved in the regulation of gene expression by glucocorticoid hormones *Biochim Biophys Acta*, *781* 1 1984

Gustafsson, J A, *et al* Structure and specific DNA binding of the rat liver glucocorticoid receptor *J Steroid Biochem*, *20* 1, 1984

Hall, J E, *et al* Control of arterial pressure and renal function during glucocorticoid excess in dogs. *Hypertension*, *2* 139, 1980

Homo-Delarche, F Glucocorticoid receptors and steroid sensitivity in normal and neoplastic human lymphoid tissues; A review *Cancer Res*, *44* 431, 1984

James, V H (ed.) The Adrenal Gland New York, Raven Press, 1979

Katzenellenbogen B S Dynamics of steroid hormone receptor action. *Annu Rev Physiol*, *42* 17, 1980

Keller Wood, M E, and Dallman, M F Corticosteroid inhibition of ACTH secretion. *Endocr Rev*, *5* 1, 1984

Kraus Friedmann, N Hormonal regulation of hepatic gluconeogenesis. *Physiol Rev*, *64* 170, 1984

Lan, N C, *et al* Mechanisms of glucocorticoid hormone action. *J Steroid Biochem*, *20* 77, 1984

Lernen, S, and Bailey, C J Thyroid hormones, gonadal and adrenocortical steroids and the function of the islets of Langerhans *Endocr Rev*, *5* 411, 1984

Miller, M Assessment of hormonal disorders of water metabolism *Clin Lab Med*, *4* 729, 1984

Morris, H G Mechanisms of action and therapeutic role of corticosteroids in asthma *J Allergy Clin Immunol*, *75* 1, 1985

Nelson, D H The Adrenal Cortex Physiological Function and Disease Philadelphia, W B Saunders Co, 1979

Rousseau G G Control of gene expression by glucocorticoid hormones *Biochem J*, *224* 1, 1984

Rousseau, G G, Structure and regulation of the glucocorticoid hormone receptor *Mol Cell Endocrinol*, *38* 1, 1984

Schmidt, T J, and Litwack, G Activation of the glucocorticoid receptor complex *Physiol Rev*, *62* 1131, 1982.

Seron Ferre, M, and Jaffe, R B. The fetal adrenal gland. *Annu Rev Physiol*, *43* 141 1981.

Sherman, M R, and Stevens, J Structure of mammalian steroid receptors Evolving concepts and methodological developments *Annu Rev Physiol*, *46* 83, 1984

Smith, P L, and McCabe, R D Mechanism and regulation of transcellular potassium transport by the colon. *Am J Physiol*, *247* G445, 1984

Young D B, and Guyton, A C: Steady state aldosterone dose-response relationships. *Circ Res*, *40(2)* 138, 1977

78

Insulin, Glucagon, and Diabetes Mellitus

The pancreas, in addition to its digestive functions, secretes two important hormones, *insulin* and *glucagon*. The purpose of this chapter is to discuss the functions of these hormones in regulating glucose, lipid, and protein metabolism, as well as to discuss briefly the two diseases—*diabetes mellitus* and *hyperinsulinism*—caused, respectively, by hyposecretion of insulin and excess secretion of insulin.

Physiologic Anatomy of the Pancreas. The pancreas is composed of two major types of tissues, as shown in Figure 78–1: (1) the *acini*, which secrete digestive juices into the duodenum, and (2) the *islets of Langerhans*, which do not have any means for emptying their secretions externally but instead secrete insulin and glucagon directly into the blood. The digestive secretions of the pancreas were discussed in Chapter 64.

The pancreas of the human being has almost a million islets of Langerhans, each only a hundred or so microns in diameter and organized around small capillaries into which its cells secrete their hormones. The islets contain three major types of cells, the *alpha*, *beta*, and *delta* cells, which are distinguished from one another by their morphology and staining characteristics. The beta cells, constituting about 60 per cent of all the cells, secrete *insulin*. The alpha cells, about 25 per cent of the total, secrete *glucagon*. And the delta cells, about 10 per cent of the total, secrete a more recently discovered hormone, *somatostatin*. In addition, at least one other type of cell, called PP cells, is present in small numbers in the islets and secretes a hormone of doubtful function called *pancreatic polypeptide*.

The close interrelationships among these different cell types in the islets of Langerhans allow direct control of secretion of some of the hormones by the other hormones. For instance, insulin inhibits glucagon secretion, and somatostatin inhibits the secretion of both insulin and glucagon.

Insulin

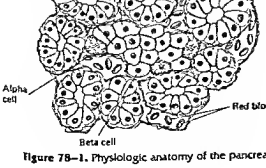


Figure 78–1. Physiologic anatomy of the pancreas.

INSULIN AND ITS METABOLIC EFFECTS

Insulin was first isolated from the pancreas in 1922 by Banting and Best, and almost overnight the outlook for the severely diabetic patient changed from one of rapid decline and death to that of a nearly normal person.

Historically, insulin has been associated with "blood sugar," and, true enough, insulin does have profound effects on carbohydrate metabolism. Yet, it is mainly abnormalities of fat metabolism, causing such conditions as acidosis and arteriosclerosis, that are the usual causes of death of a diabetic patient. And, in patients with prolonged diabetes, the inability to synthesize proteins leads to wasting of the tissues as well as many cellular functional disorders. Therefore, it is clear that insulin affects fat and protein metabolism almost as much as it does carbohydrate metabolism.

Chemistry of Insulin. Insulin is a small protein with a molecular weight of 5808 for human insulin. It is composed of two amino acid chains, illustrated in Figure 78–2, connected to each other by disulfide linkages. When the two amino acid chains are split apart, the functional activity of the insulin molecule is lost.

Insulin is synthesized in the beta cells by the usual cell machinery for protein synthesis as explained in Chapter 3, beginning with translation of the insulin RNA by ribosomes attached to the endoplasmic reticu-

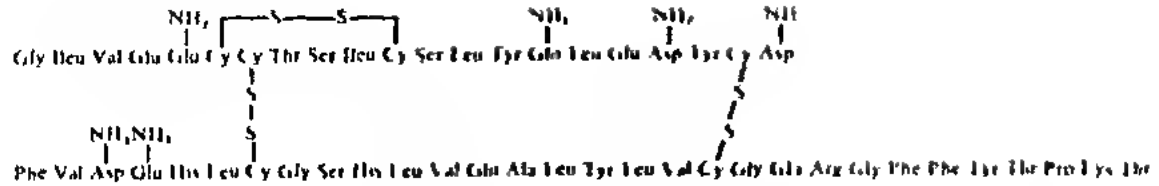


Figure 78–2. The human insulin molecule

lum to form an *insulin preprohormone*. This initial preprohormone has a molecular weight of about 11,500, but it is then cleaved in the endoplasmic reticulum to form a *proinsulin* with a molecular weight of about 9000, most of this is further cleaved in the Golgi apparatus to form insulin before being packaged in the secretory granules. However, about one sixth of the final secreted product is still in the form of proinsulin. Unfortunately, the proinsulin has no insulin activity.

When insulin is secreted into the blood, it circulates almost entirely in an unbound form, it has a plasma half-life averaging only about 6 minutes so that it is mainly cleared from the circulation within 10 to 15 minutes. Except for that portion of the insulin that combines with receptors in the target cells, the remainder is mainly degraded in the liver and to a lesser extent in the kidneys. This rapid removal from the plasma is important because at times it is equally as important to rapidly turn off as to turn on the control functions of insulin.

To initiate its effects in target cells, insulin first binds with a membrane receptor protein having a molecular weight of about 300,000. This activates the receptor that initiates the insulin effects, but from here the cellular molecular mechanisms are vague. The activated receptors do mildly activate the cyclic AMP system of the cell, which is believed then to act as a second messenger to promote some of the insulin effects. However, most insulin effects occur in the absence of increased cyclic AMP, thus suggesting other intermediary processes.

EFFECT OF INSULIN ON CARBOHYDRATE METABOLISM

Immediately after a high carbohydrate meal, the glucose that is absorbed into the blood causes rapid secretion of insulin, which we shall discuss in detail later in the chapter. The insulin in turn causes rapid uptake, storage, and use of glucose by almost all tissues of the body, but especially by the liver, muscles, and adipose tissue. Therefore, let us discuss each of these.

Effect of Insulin on Promoting Liver Uptake, Storage, and Use of Glucose

One of the most important of all the effects of insulin is to cause most of the glucose absorbed after a meal to be stored almost immediately in the liver in the form of glycogen. Then, between meals, when food is not available and the blood glucose concentration begins to fall, the liver glycogen is split back into glucose, which is released back into the blood to keep the blood glucose concentration from falling too low.

The mechanism by which insulin causes glucose uptake and storage in the liver includes several almost simultaneous steps:

(1) Insulin *inhibits phosphorylase*, the enzyme that causes liver glycogen to split into glucose. This obviously prevents breakdown of the glycogen that is already in the liver cells.

(2) Insulin causes *enhanced uptake of glucose* from the blood by the liver cells. It does this by *increasing the activity of the enzyme glucokinase*, which is the enzyme that causes the initial phosphorylation of glucose after it diffuses into the liver cells. Once phosphorylated, the glucose is trapped inside the liver cells, because phosphorylated glucose cannot diffuse back through the cell membrane.

(3) Insulin also increases the activities of the enzymes that promote glycogen synthesis, including *phosphofructokinase* that causes the second stage in the phosphorylation of the glucose molecule and *glycogen synthetase* that is responsible for polymerization of the monosaccharide units to form the glycogen molecules.

The net effect of all these actions is to increase the amount of glycogen in the liver. The glycogen can increase to a total of about 5 to 6 per cent of the liver mass, which is equivalent to almost 100 grams of stored glycogen.

Release of Glycogen from the Liver Between Meals. After the meal is over and the blood glucose level begins to fall to a low level, several events now transpire that cause the liver to release glucose back into the circulating blood:

1. The decreasing blood glucose causes the pancreas to decrease its insulin secretion.
2. The lack of insulin then reverses all the effects listed above for glycogen storage, essentially stopping further synthesis of glycogen in the liver. This also prevents further uptake of glucose by the liver from the blood.
3. The lack of insulin (along with increase of glucagon, which will be discussed later) activates the enzyme *phosphorylase* which causes the splitting of glycogen into *glucose phosphate*.
4. The enzyme *glucose phosphatase* causes the phosphate radical to split away from the glucose, and this allows the free glucose to diffuse back into the blood.

Thus, the liver removes glucose from the blood when it is present in excess after a meal and returns it to the blood when it is needed between meals. Ordinarily, about 60 per cent of the glucose in the meal is stored in this way in the liver and then returned later.

Other Effects of Insulin on Carbohydrate Metabolism in the Liver. Insulin also *promotes the conversion of excess liver glucose into fatty acids*, and these fatty acids are subsequently transported to the adipose tissue and deposited as fat. Insulin also *inhibits gluconeogenesis*. It does this mainly by decreasing the quantities and activities of the liver enzymes required for gluconeogenesis. However, part of the effect is caused by an action of insulin to decrease the release of amino acids from muscle and other extrahepatic tissues, thus decreasing the availability of the necessary precursors required for gluconeogenesis This will be discussed further in relation to the effect of insulin on protein metabolism.

Effect of Insulin in Promoting Glucose Metabolism in Muscle

During most of the day, muscle tissue depends not on glucose for its energy but instead on fatty acids. The principal reason for this is that the normal *resting muscle* membrane is almost impermeable to glucose except when the muscle fiber is stimulated by insulin. And, between meals, the amount of insulin that is secreted is too small to promote significant amounts of glucose entry into the muscle cells

However, under two conditions the muscles do utilize large amounts of glucose. One of these is during periods of heavy exercise. This usage of glucose does not require large amounts of insulin because exercising muscle fibers, for reasons not understood, become highly permeable to glucose even in the absence of insulin because of the contraction process itself.

The second condition for muscle usage of large amounts of glucose is during the few hours after a meal. At this time the blood glucose concentration is high; also, the pancreas is secreting large quantities of insulin, and the extra insulin causes rapid transport of glucose into the muscle cells. This causes the muscle cell during this period of time to utilize carbohydrates preferentially over fatty acids.

Storage of Glycogen in Muscle. If the muscles are not exercising during the period following a meal and yet glucose is transported into the muscle cells in abundance, then most of the glucose is stored in the form of muscle glycogen instead of being used for energy. However, the concentration of muscle glycogen rarely rises much above 1 to 2 per cent rather than the possible 5 to 6 per cent in liver cells. The glycogen can later be used for energy by the muscle. This glycogen is especially useful to provide extreme spurts of anaerobic energy for a few minutes at a time by glycolytic breakdown of the glycogen to lactic acid, which can occur in the absence of oxygen.

Muscle glycogen is different from liver glycogen in that it cannot be reconverted into free glucose and released back into the blood because there is no glucose phosphatase in muscle cells, in contrast to the liver cells.

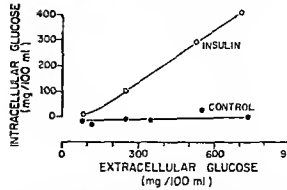


Figure 78–3. Effect of insulin in enhancing the concentration of glucose inside muscle cells. Note that in the absence of insulin (control) the intracellular glucose concentration remained near zero despite very high extracellular glucose concentrations of glucose (from Park, Morgan, Kajl, and Smith, *in* Eisenstein [ed.] The Biochemical Aspects of Hormone Action. Little, Brown and Co)

Insulin Facilitation of Glucose Transport Through the Muscle Cell Membrane. Insulin promotes glucose transport into muscle cells quite differently from the way it promotes transport into liver cells. The transport into liver results mainly from a trapping mechanism caused by phosphorylation of the glucose under the influence of glucokinase. However, this is only a minor factor in the insulin effect on glucose transport into muscle cells. Of more importance, the insulin has a direct effect on the muscle cell membrane to *facilitate glucose transport*. This is illustrated by the experimental results depicted in Figure 78–3. This experiment was performed in muscle cells at a temperature of 4° C, a temperature at which glucose, upon entering the cell, cannot be phosphorylated. The lower curve labeled "control" shows the concentration of free glucose measured inside the cell, illustrating that the glucose concentration remained almost exactly zero despite increases in extracellular glucose concentration up to as high as 750 mg/dl. In contrast, the curve labeled "insulin" illustrates that the intracellular glucose concentration rose to as high as 400 mg when insulin was added. Thus, it is clear that insulin can increase the rate of transport of glucose into the resting muscle cell by at least 15- to 20-fold.

As pointed out in the discussion of glucose transport through the cell membrane in Chapter 9, glucose cannot pass through the usual membrane pores but instead must be transported by a specific "carrier" protein. Figure 78–4 depicts the generally accepted method by which this is achieved, showing that when the carrier is activated by insulin, glucose combines with the carrier and diffuses in association with the carrier to the inside of the membrane. Then the glucose is released to the interior of the cell. The carrier is used again and again to transport additional q[illegible] glucose. This process is called *facilitated d[illegible]*

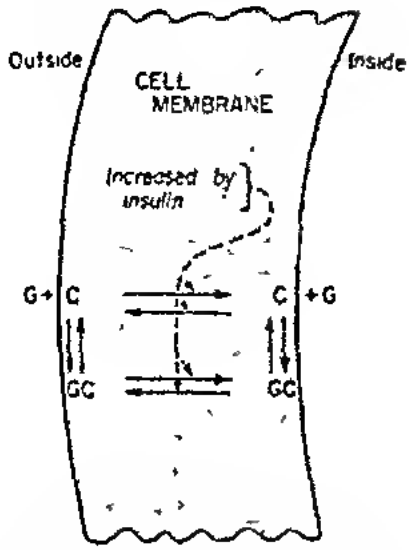


Figure 78–4. Effect of insulin in increasing glucose transport in either direction through the cell membrane

and it can occur in *either direction* as shown by the reversible arrows in the diagram

The manner in which insulin enhances facilitated diffusion of glucose is still largely unknown All that is known is that the insulin combines with a "receptor" protein in the cell membrane—a protein having a molecular weight of about 300,000 This may be the glucose carrier itself, or it may be merely the first step in a chain of events that leads to activation of the carrier system. The insulin increases the glucose transport within seconds, suggesting either a rapid direct action in the cell membrane itself or some other equally rapid mechanism

Lack of Effect of Insulin on Glucose Uptake and Usage by the Brain

The brain is quite different from most other tissues of the body in that insulin has either little or no effect on uptake or use of glucose. Instead, *the brain cells are permeable to glucose without the intermediation of insulin.*

The brain cells normally use only glucose for energy Therefore, it is essential that the blood glucose level be maintained always above a critical level, which is one of the important functions of the blood glucose control system. When the blood glucose does fall too low, into the range of 20 to 50 mg/dl, symptoms of *hypoglycemic shock* develop, characterized by progressive irritability that leads to fainting, convulsions, and even coma.

Effect of Insulin on Carbohydrate Metabolism in Other Cells

Insulin increases glucose transport into and glucose usage by most other cells of the body (with the exception of the liver and brain cells, as noted) the same way that it affects glucose transport through the muscle cell membrane. The transport of glucose into adipose cells is essential for deposition of fat in these cells, as will be discussed next.

EFFECT OF INSULIN ON FAT METABOLISM

Though not quite as dramatic as the acute effects of insulin on carbohydrate metabolism, insulin also affects fat metabolism in ways that, in the long run, are perhaps equally as important. Especially dramatic is the long-term effect of insulin lack in causing extreme atherosclerosis, often leading to heart attacks, cerebral strokes, and other vascular accidents. But, first, let us discuss the acute effects of insulin on fat metabolism.

Effect of Insulin Excess on Fat Synthesis and Storage

Insulin has several different effects that lead to fat storage in adipose tissue One is the simple fact that insulin increases the utilization of glucose by most of the body's tissues, which automatically decreases the utilization of fat, thus functioning as a "fat sparer" However, insulin also promotes fatty acid synthesis. Most of this synthesis occurs in the liver cells, and the fatty acids are then transported to the adipose cells to be stored. However, a small part of the synthesis occurs in the fat cells themselves. The different factors that lead to increased fatty acid synthesis in the liver include:

1. Insulin increases the transport of glucose into the liver cells After the liver glycogen concentration increases to 5 to 6 per cent, this in itself inhibits further glycogen synthesis. Then, all the additional glucose entering the liver cells becomes available to form fat. The glucose is first split to pyruvate in the glycolytic pathway, and the pyruvate subsequently is converted to acetyl-CoA, the substrate from which fatty acids are synthesized.

2. An excess of *citrate* and *isocitrate ions* is formed by the citric acid cycle when excess amounts of glucose are being used for energy. These ions then have a direct effect in activating *acetyl-CoA carboxylase*, the enzyme required to initiate the first stage of fatty acid synthesis.

3. Fatty acids are then transported in the blood from the liver to the adipose cells where they are stored.

4 Large portions of the fatty acids are also utilized within the liver itself to synthesize triglycerides, the usual form of storage fat. Most of these are then released from the liver cells to the blood in the lipoproteins. Insulin activates lipoprotein lipase in the adipose tissue, which splits the triglycerides again into fatty acids, a requirement for them to be absorbed into the adipose cells where they are again converted to triglycerides and stored.

Storage of Fat in the Adipose Cells. Insulin

has very much the same effect in adipose cells as in the liver in causing synthesis of fatty acids. However, only about one tenth as much glucose is transported into the fat cells of the human being as into the liver, so that the amount of fatty acids synthesized in adipose cells is rather small compared with the amount formed in the liver.

Yet, insulin has two other essential effects that are required for fat storage in adipose cells:

1. Insulin *inhibits the action of hormone-sensitive lipase*. Since this is the enzyme that causes hydrolysis of the triglycerides in fat cells, the release of fatty acids into the circulating blood is therefore inhibited.

2. Insulin *promotes glucose transport into the fat cells* in exactly the same way that it promotes glucose transport into muscle cells. Some of this glucose is then utilized to synthesize fatty acids, as noted earlier, but, more important, it also forms another substance that is essential to the storage of fat. During the glycolytic breakdown of glucose, large quantities of the substance *α-glycerophosphate* is formed. This substance supplies the *glycerol* that binds with fatty acids to form the triglycerides that are the storage form of fat in adipose cells. Therefore, when insulin is not available to promote glucose entry into the fat cells, fat storage is almost totally blocked

Increased Metabolic Use of Fat Caused by Insulin Lack

All aspects of fat metabolism are greatly *enhanced in the absence of insulin*. This occurs even normally between meals when secretion of insulin is minimal, but it becomes extreme in diabetes when secretion of insulin is almost zero. The resulting effects are:

Lipolysis of Storage Fat and Release of Free Fatty Acids During Insulin Lack. In the absence of insulin, all the effects of insulin noted above causing storage of fat are reversed. The most important effect is that the enzyme *hormone-sensitive lipase* in the fat cells becomes strongly activated. This causes hydrolysis of the stored triglycerides, releasing large quantities of fatty acids and glycerol into the circulating blood. Consequently, the plasma concentration of free fatty acids rises within minutes to hours. This free fatty acid then becomes the main energy substrate used by essentially all tissues of the body besides the brain. Figure 78–5 illustrates the effect of insulin lack on the plasma concentrations of free fatty acids, glucose, and acetoacetic acid. Note that immediately after removal of the pancreas the free fatty acid concentration in the plasma begins to rise, rising considerably more rapidly even than the concentration of glucose.

Effect of Insulin Lack on Causing a Fatty Liver. Strange as it may seem, though insulin lack causes dissolution of the fat stores in adipose

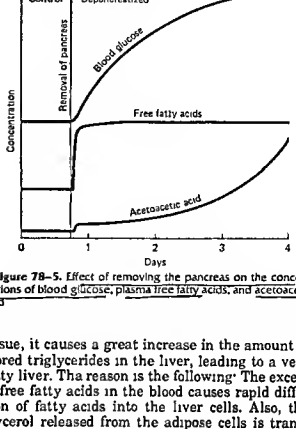


Figure 78–5. Effect of removing the pancreas on the concentrations of blood glucose, plasma free fatty acids, and acetoacetic acid

tissue, it causes a great increase in the amount of stored triglycerides in the liver, leading to a very fatty liver. The reason is the following: The excess of free fatty acids in the blood causes rapid diffusion of fatty acids into the liver cells. Also, the glycerol released from the adipose cells is transported at the same time into the liver. The liver cells, unlike the adipose cells, have appropriate enzymes for converting glycerol into α-glycerophosphate, and this causes rapid binding of the fatty acids to form triglycerides. Thus, insulin lack causes fatty acids and glycerol to leave the adipose cells, but the liver captures a large share of the released fatty acids and glycerol and reconverts these to triglycerides that are mainly stored in the liver. In this manner, the liver can store fat in amounts equal to 30 per cent or more of its weight.

Effect of Insulin Lack on Plasma Lipid Concentrations. The excess of fatty acids in the liver also promotes conversion of some of the fatty acids into phospholipids and cholesterol, two of the major products of fat metabolism. These two substances, along with some of the triglycerides formed in the liver, are then discharged into the blood in the lipoproteins. Occasionally, the plasma lipoproteins increase as much as threefold in the absence of insulin, giving a total concentration of plasma lipids of several per cent rather than the normal 0.6 per cent. This high lipid concentration—especially the high concentration of cholesterol—leads to rapid development of atherosclerosis in persons with serious diabetes.

Ketogenic and Acidotic Effects of Insulin Lack. Insulin lack also causes excessive amounts of *acetoacetic acid* to be formed in the liver cells. This results from the following effect: In the absence of insulin but in the presence of excess fatty acids in the liver cells, the carnitine transport

mechanism for transporting fatty acids into the mitochondria becomes increasingly activated. In the mitochondria, beta oxidation of the fatty acids then proceeds exceedingly rapidly, releasing extreme amounts of acetyl-CoA. A part of this acetyl-CoA can be utilized in the liver itself for energy, but the excess is condensed to form acetoacetic acid, which, in turn, is released into the circulating blood. Most of this passes to the peripheral cells where it is again converted into acetyl-CoA and used for energy in the usual manner.

However, in the absence of insulin, the utilization of acetoacetic acid in the peripheral tissues is depressed, while at the same time so much acetoacetic acid is released from the liver that it cannot all be metabolized by the tissues. Therefore, as illustrated in Figure 78–5, its concentration rises during the days following cessation of insulin secretion, sometimes reaching a concentration as high as 10 or more milliequivalents per liter. As was explained in Chapter 68, some of the acetoacetic acid is also converted into β-hydroxybutyric acid and *acetone*. These two substances, along with the acetoacetic acid, are called *ketone bodies*, and their presence in large quantities in the body fluids is called *ketosis*. We shall see later that the acetoacetic acid and the β-hydroxybutyric acid can cause severe *acidosis* and *coma* in patients with severe diabetes, which often leads to death.

Effect of Excess Plasma Fatty Acids on Inhibiting Glucose Utilization by Cells. It is already clear that lack of insulin decreases glucose uptake by most cells of the body and in this way greatly depresses glucose utilization by the cells when insulin is not available. However, still another effect occurs that further depresses glucose utilization. The increased fatty acids in the plasma caused by insulin lack have an additional direct effect on the cells themselves to depress glucose utilization. This results from the following mechanisms:

First, when the fatty acids enter the cells, they are immediately split to form acetyl-CoA. This then enters the citric acid cycle and provides the energy required by the cell.

Second, two of the important products of this energy system are *citrate ion* and *ATP*, both of which have a strong inhibitory effect on the enzyme *phosphofructokinase*, which is a rate-limiting enzyme for promoting glucose use in the cell. Therefore, use of glucose for energy almost ceases.

Third, since glucose cannot be utilized by the cell for energy, the mass law effect causes glucose uptake into the cell also to be greatly depressed.

This total process is called the *glucose–fatty acid cycle*. Its net effect is approximately to double the depressive action of insulin lack on cellular uptake and utilization of glucose. This obviously is important to prevent utilization of glucose by the cells of the body besides those of the brain when glucose is in short supply.

EFFECT OF INSULIN ON PROTEIN METABOLISM AND GROWTH

Effect of Insulin on Protein Synthesis and Storage. During the few hours following a meal when excess quantities of nutrients are available in the circulating blood, not only carbohydrates and fats but proteins as well are stored in the tissues; insulin is required for this to occur. The manner in which insulin causes protein storage is not as well understood as the mechanisms for both glucose and fat storage. Some of the facts known are:

1. Insulin causes active transport of many of the amino acids into the cells. Among the amino acids most strongly transported are *valine, leucine, isoleucine, tyrosine,* and *phenylalanine*. Thus, insulin shares with growth hormone the capability of increasing the uptake of amino acids into cells. However, the amino acids affected are not necessarily the same ones.

2. Insulin has a direct effect on the ribosomes to *increase the translation of messenger RNA*, thus forming new proteins. In some unexplained way, insulin "turns on" the ribosomal machinery. In the absence of insulin the ribosomes simply stop working, almost as if insulin operates an "on-off" mechanism.

3. Over a longer period of time insulin also *increases the rate of transcription of DNA* in the cell nuclei, thus forming increased quantities of RNA. Eventually, it also increases the rate of formation of new DNA and thus promotes reproduction of cells. All these effects promote still more protein synthesis.

4. Insulin also *inhibits the catabolism of proteins*, thus decreasing the rate of amino acid release from the cells, especially from the muscle cells. Presumably this results from some ability of the insulin to diminish the normal degradation of proteins by the cellular lysosomes.

5. In the liver, insulin *depresses the rate of gluconeogenesis*. It does this by decreasing the activity of the enzymes that promote gluconeogenesis. Since the substrates most used for synthesis of glucose by the process of gluconeogenesis are the plasma amino acids, this suppression of gluconeogenesis conserves the amino acids in the protein stores of the body.

In summary, insulin promotes protein formation and also prevents the degradation of proteins. However, we still know very little about the basic chemical mechanisms employed by insulin to achieve these ends.

Protein Depletion and Increased Plasma Amino Acids Caused by Insulin Lack. All protein storage comes to a complete halt when insulin is not available. The catabolism of proteins increases, protein synthesis stops, and large quantities of amino acid are dumped into the plasma. The plasma amino acid concentration rises consid-

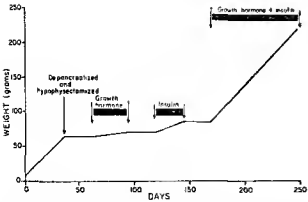


Figure 78–6. Effect of (a) growth hormone, (b) insulin, and (c) growth hormone plus insulin on growth in a depancreatized and hypophysectomized rat

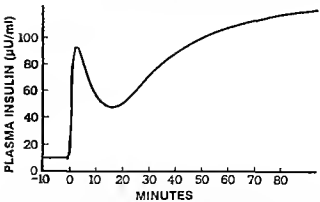


Figure 78–7. Increase in plasma insulin concentration following a sudden increase in blood glucose to two to three times the normal range. Note an initial rapid surge in insulin concentration and then a delayed but higher and continuing increase in concentration beginning 15 to 20 minutes later.

erably, and most of the excess amino acids are either used directly for energy or as substrates for gluconeogenesis. This degradation of the amino acids also leads to enhanced urea excretion in the urine. The resulting protein wasting is one of the most serious of all the effects of severe diabetes mellitus. It can lead to extreme weakness as well as to many deranged functions of the organs.

Effect of Insulin on Growth—Its Synergistic Effect with Growth Hormone. Because insulin is required for the synthesis of proteins, it is equally as essential for growth of an animal as is growth hormone. This is illustrated in Figure 78–6, which shows that a depancreatized hypophysectomized rat without therapy hardly grows at all. Furthermore, administration of either growth hormone or insulin one at a time causes almost no growth. Yet a combination of both these hormones does cause dramatic growth. Thus it appears that the two hormones function synergistically to promote growth, each performing its specific function that is separate from the other. Perhaps part of this necessity for both hormones results from the fact that each promotes cellular uptake of a different selection of amino acids, all of which are required if growth is to be achieved.

CONTROL OF INSULIN SECRETION

Formerly, it was believed that insulin secretion is controlled almost entirely by the blood glucose concentration. However, as more has been learned about the metabolic functions of insulin for protein and fat metabolism, it has been learned that blood amino acids and other factors also play important roles in controlling insulin secretion.

Stimulation of Insulin Secretion by Blood Glucose. At the normal fasting level of blood glucose of 80 to 90 mg/dl, the rate of insulin secretion is minimal—in the order of 25 ng/min/kg (600 μUnits/min/kg) of body weight. If the blood glucose concentration is suddenly increased to a level two to three times normal and is kept at this high level thereafter, insulin secretion increases markedly in two stages, as shown in Figure 78–7:

1. Insulin secretion increases almost tenfold within 3 to 5 minutes after acute elevation of the blood glucose; this results from immediate dumping of preformed insulin from the beta cells of the islets of Langerhans. However, this initial high rate of secretion is not maintained; instead, it decreases about halfway back toward normal in another 5 to 10 minutes.

2. After about 15 minutes, insulin secretion rises a second time, reaching a new plateau in 2 to 3 hours, this time usually at a rate of secretion even greater than that in the initial phase. This secretion results both from additional release of preformed insulin and from activation of some enzyme system that synthesizes and releases new insulin from the cells.

Relationship Between Blood Glucose Concentration and Insulin Secretion Rate. As the concentration of blood glucose rises above 100 mg/dl of blood, the rate of insulin secretion rises rapidly, reaching a peak some 10 to 30 times the basal level at blood glucose concentrations between 400 and 600 mg/dl, as illustrated in Figure 78–8. Thus, the increase in insulin secretion under a glucose stimulus is dramatic both in its rapidity and in the tremendous level of secretion achieved.

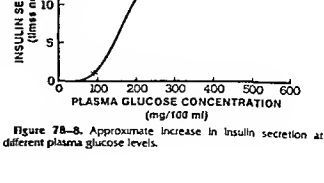


Figure 78–8. Approximate increase in insulin secretion at different plasma glucose levels.

Furthermore, the turn-off of insulin secretion is almost equally as rapid, occurring within minutes after reduction in blood glucose concentration back to the fasting level.

This response of insulin secretion to an elevated blood glucose concentration provides an extremely important feedback mechanism for regulating blood glucose concentration. That is, the rise in blood glucose increases insulin secretion, and the insulin in turn causes transport of glucose into the liver, muscle, and other cells, thereby reducing the blood glucose concentration toward the normal value.

Other Factors That Stimulate Insulin Secretion

Amino Acids. In addition to excess glucose stimulating insulin secretion, many of the amino acids have a similar effect. The most potent of these are *arginine* and *lysine*. However, this effect differs from glucose stimulation of insulin secretion in the following way: Amino acids administered in the absence of a rise in blood glucose cause only a small increase in insulin secretion. However, when administered at the same time that the blood glucose concentration is elevated, the glucose-induced secretion of insulin may be as much as doubled in the presence of the excess amino acids. Thus, the amino acids very strongly potentiate the glucose stimulus for insulin secretion.

The stimulation of insulin secretion by amino acids seems to be a purposeful response because the insulin in turn promotes transport of the amino acids into the tissue cells and also promotes intracellular formation of protein. That is, the insulin is important for proper utilization of excess amino acids in the same way that it is important for the utilization of carbohydrates.

Gastrointestinal Hormones. A mixture of several important gastrointestinal hormones—*gastrin*, *secretin*, *cholecystokinin*, and *gastric inhibitory peptide* (which seems to be the most potent of them all)—will cause a moderate increase in insulin secretion. These hormones are released in the gastrointestinal tract after a person eats a meal. They then cause an "anticipatory" increase in blood insulin in preparation for the glucose and amino acids to be absorbed from the meal. These gastrointestinal hormones almost double the rate of insulin secretion following an average meal.

Other Hormones and the Autonomic Nervous System. Other hormones that either directly increase insulin secretion or potentiate the glucose stimulus for insulin secretion include *glucagon*, *growth hormone*, *cortisol*, and to a lesser extent *progesterone* and *estrogen*. The importance of the stimulatory effects of these hormones is that prolonged secretion of any one of them in large quantities can occasionally lead to exhaustion of the beta cells of the islets of Langerhans and thereby cause diabetes. Indeed, diabetes often occurs in persons maintained on high pharmacological doses of some of these hormones. It is particularly common in giants or acromegalic persons with growth hormone–secreting tumors or in persons who already have a diabetic tendency to whom glucocorticoids are administered in high concentration.

Under some conditions, stimulation of either the parasympathetic or the sympathetic nerves to the pancreas can also increase insulin secretion. However, it is doubtful that either of these effects is of physiological significance for regulating insulin secretion.

ROLE OF INSULIN IN "SWITCHING" BETWEEN CARBOHYDRATE AND LIPID METABOLISM

From the preceding discussions it should already be clear that insulin promotes the utilization of carbohydrates for energy, while it depresses the utilization of fats. Conversely, lack of insulin causes fat utilization mainly to the exclusion of glucose utilization, except by brain tissue. Furthermore, the signal that controls this switching mechanism is principally the blood glucose concentration. When the glucose concentration is low, insulin secretion is suppressed and fat is utilized almost exclusively for energy everywhere except in the brain; when the glucose concentration is high, insulin secretion is stimulated, and carbohydrate is utilized instead of fat until the excess blood glucose is stored in the form of liver glycogen or muscle glycogen. Therefore, one of the most important functional roles of insulin in the body is to control which of these two foods from moment to moment will be utilized by the cells for energy.

At least four other known hormones also play important roles in this switching mechanism: *growth hormone* from the anterior pituitary gland, *cortisol* from the adrenal cortex, *epinephrine* from the adrenal medulla, and *glucagon* from the alpha cells of the islets of Langerhans in the pancreas. Glucagon will be discussed in the following section of this chapter. Both growth hormone and cortisol are secreted in response to hypoglycemia, and both inhibit cellular utilization of glucose while promoting fat utilization; however, the effects of both these hormones develop very slowly, requiring 30 minutes to several hours to become maximal.

Epinephrine is especially important in increasing plasma glucose concentration during periods of stress when the sympathetic nervous system is excited. However, epinephrine acts differently from the other hormones in that it increases the plasma fatty acid concentration at the same time. The reasons for these effects are: (1) Epinephrine has the very potent effect of causing glycogenolysis in the liver, thus releasing large quantities of glucose into the blood, and (2) it also has a direct lipolytic effect on the adipose cells because it activates a tissue lipase, thus greatly enhancing the blood concentration of fatty acids as well. Quantitatively, the enhancement of fatty acids is far greater than the enhancement of blood glucose. Therefore, epinephrine especially enhances the utilization of fat in such stressful states as exercise, circulatory shock, anxiety, and so forth.

GLUCAGON AND ITS FUNCTIONS

Glucagon, a hormone secreted by the alpha cells of the islets of Langerhans when the blood glucose

concentration falls, has several functions that are diametrically opposed to those of insulin. Most important of these is to increase the blood glucose concentration, an effect which is exactly opposite to that of insulin.

Like insulin, glucagon is a large polypeptide. It has a molecular weight of 3485 and is composed of a chain of 29 amino acids. An injection of purified glucagon into an animal, a profound *hyper*glycemic effect occurs. One microgram per kilogram of glucagon can elevate the blood glucose concentration approximately 20 mg/dl of blood in about 20 minutes. For this reason, glucagon is also called *hyperglycemic factor*.

EFFECTS ON GLUCOSE METABOLISM

The two major effects of glucagon on glucose metabolism are (1) breakdown of liver glycogen (*glycogenolysis*) and (2) increased *gluconeogenesis* in the liver. Both these effects greatly enhance the availability of glucose to the other organs of the body.

Glycogenolysis and Increased Blood Glucose Concentration Caused by Glucagon. The most dramatic effect of glucagon is its ability to cause glycogenolysis in the liver, which in turn increases the blood glucose concentration within minutes.

It does this by the following complex cascade of events:

1. Glucagon activates *adenyl cyclase* in the hepatic cell membrane,
2. Which causes the formation of *cyclic AMP*,
3. Which activates *protein kinase regulator protein*,
4. Which activates *protein kinase*,
5. Which activates *phosphorylase b kinase*,
6. Which converts *phosphorylase b* into *phosphorylase a*,
7. Which promotes the degradation of glycogen into glucose-1-phosphate,
8. Which then is dephosphorylated and the glucose released from the liver cells.

This sequence of events is exceedingly important for several reasons. First, it is one of the most thoroughly studied of all the *second messenger* functions of cyclic AMP. Second, it illustrates a cascading system in which *each succeeding product is produced in greater quantity than the preceding product*. Therefore, it represents a potent *amplifying* mechanism. This explains how only 1 μg/kg of glucagon can have the extreme effect of causing hyperglycemia.

Infusion of glucagon for about four hours can cause such intensive liver glycogenolysis that all the liver stores of glycogen become totally depleted.

Gluconeogenesis Caused by Glucagon. Even after all the glycogen in the liver has been exhausted under the influence of glucagon, continued infusion of this hormone still causes continued hyperglycemia. This results from an effect of glucagon in increasing the rate of gluconeogenesis in the liver cells. Unfortunately, the precise mechanism of the stimulatory effect on gluconeogenesis is unknown, but it is believed to be activation of the enzyme system for converting pyruvate to phosphoenolpyruvate, a rate-limiting step in gluconeogenesis. Also, glucagon increases the extraction of amino acids from the blood by the liver cells, thus making a greater quantity of these available to be converted into glucose.

Other Effects of Glucagon

Most other effects of glucagon occur only when its concentration rises well above that normally in the blood. Perhaps the most important effect is that glucagon mildly activates adipose cell lipase, making increased quantities of fatty acids available to the energy systems of the body. It also inhibits the storage of triglycerides in the liver, which prevents the liver from removing fatty acids from the blood; this also helps make additional amounts of fatty acids available for the other tissues of the body.

Glucagon in very large concentrations also (a) enhances the strength of the heart, (b) enhances bile secretion, and (c) inhibits gastric acid secretion. However, all these changes are probably unimportant in the normal function of the body.

REGULATION OF GLUCAGON SECRETION

Effect of Blood Glucose Concentration. The blood glucose concentration is by far the most potent factor controlling glucagon secretion. Note specifically, however, that the effect of blood glucose concentration on glucagon secretion is in exactly the opposite direction from the effect of glucose on insulin secretion. This is illustrated in Figure 78–9, showing that a decrease in the blood glucose concentration from its normal fasting level of about 90 mg/dl of blood down to hypoglycemic levels can increase glucagon secretion as well as its plasma concentration severalfold. On the other hand, increasing the blood glucose to hyperglycemic levels decreases plasma glucagon. Thus, in hypoglycemia, glucagon is secreted in large

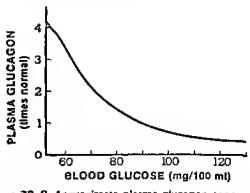


Figure 78–9. Approximate plasma glucagon concentration at different blood glucose levels

amounts; it then mobilizes glucose from the liver and thereby helps correct the hypoglycemia.

Effect of Amino Acids. High concentrations of amino acids, as occur in the blood after a protein meal (especially the amino acids *alanine* and *arginine*), stimulate the secretion of glucagon. This is the same effect that amino acids have in stimulating insulin secretion. Thus, in this instance the glucagon and insulin responses are not opposite to each other.

The importance of amino acid stimulation of glucagon secretion is believed to be the following: After a meal, amino acids stimulate insulin secretion, and the insulin in turn causes deposition of amino acids in cells throughout the body. However, the insulin also tends to cause hypoglycemia at the same time. This is partly or totally offset by the simultaneous increase in glucagon secretion caused by the blood amino acids because the glucagon mobilizes glucose from the liver. This glucose also provides additional cellular energy, which is another requirement for synthesis of proteins in the tissues.

Effect of Exercise. In exhaustive exercise the blood concentration of glucagon often increases as much as four- to fivefold. However, what causes this is not understood because the blood glucose concentration does not necessarily fall. Nevertheless, a beneficial effect of the glucagon is that it prevents a decrease in blood glucose. One of the factors that might increase glucagon secretion in exercise is increased circulating amino acids. However, other factors, such as nervous stimulation of the islets of Langerhans, could also play a role.

SOMATOSTATIN—ITS EFFECT ON INHIBITION OF GLUCAGON AND INSULIN SECRETION

The *delta cells* of the islets of Langerhans secrete the hormone *somatostatin*, which is a polypeptide containing only 14 amino acids that has the extremely short half-life in the circulating blood of only two minutes. Almost all factors related to the ingestion of food will stimulate somatostatin secretion. These include (1) increased blood glucose, (2) increased amino acids, (3) increased fatty acids, and (4) increased concentrations of several of the gastrointestinal hormones released from the upper gastrointestinal tract in response to food intake.

In turn, somatostatin has multiple inhibitory effects as follows.

(1) Somatostatin acts locally within the islets of Langerhans themselves to depress the secretion of both insulin and glucagon.

(2) Somatostatin decreases the motility of the stomach, the duodenum, and the gallbladder.

(3) Somatostatin decreases both secretion and absorption in the gastrointestinal tract.

Putting all this information together, it has been suggested that the principal role of somatostatin is to cause slowing of the assimilation of food from the gut. That is, following ingestion of a meal multiple factors stimulate somatostatin secretion; the somatostatin then extends the period of time over which the food nutrients are assimilated into the blood. At the same time, the effect of somatostatin to depress insulin and glucagon secretion decreases the utilization of the absorbed nutrients by the tissues, thus preventing rapid exhaustion of the food and therefore making it available over a longer period of time.

It should also be recalled that somatostatin is the same chemical substance as *growth hormone inhibitory hormone* that is secreted in the hypothalamus and is transported to the anterior pituitary gland to suppress growth hormone secretion.

PANCREATIC POLYPEPTIDE

The islets of Langerhans also secrete a fourth hormone, *pancreatic polypeptide*, secreted by the *PP cells*. Unfortunately, the functions of this hormone are mainly unknown, though recent experiments suggest that it might play a role in controlling certain gastrointestinal functions. Pancreatic polypeptide is mentioned here merely to emphasize that many functions of the islets of Langerhans are yet to be discovered.

SUMMARY OF BLOOD GLUCOSE REGULATION

In the normal person the blood glucose concentration is very narrowly controlled, usually in a range between 80 and 90 mg/dl of blood in the fasting person each morning before breakfast. This concentration increases to 120 to 140 mg/dl during the first hour or so following a meal, but the feedback systems for control of blood glucose return the glucose concentration very rapidly back to the control level, usually within two hours after the last absorption of carbohydrates. Conversely, in starvation the gluconeogenesis function of the liver provides the glucose that is required to maintain the fasting blood glucose level.

The mechanisms for achieving this high degree of control have been presented in this chapter. Let us summarize these briefly:

1. The liver functions as a very important *blood glucose–buffer system*. That is, when the blood glucose rises to a very high concentration following a meal and the rate of insulin secretion also increases, as much as two thirds of the glucose absorbed from the gut is almost immediately stored in the liver in the form of glycogen. Then, during the succeeding hours, when both the blood glucose concentration and the rate of insulin secretion fall, the liver releases the glucose back into the blood. In this way, the liver decreases the variations in blood glucose concentration by about threefold. In fact, in patients with severe liver disease, it becomes almost impossible to maintain a narrow range of blood glucose concentration.

2. It is very clear that both insulin and glucagon function as important and separate feedback control systems for maintaining a normal blood glucose concentration. When the concentration rises

to a level too high, insulin is secreted; the insulin in turn causes the blood glucose concentration to decrease toward normal. Conversely, a decrease in blood glucose stimulates glucagon secretion; the glucagon then functions in the opposite direction to increase the glucose up toward normal. Under most normal conditions, the insulin feedback mechanism is much more important than the glucagon mechanism, but in instances of diminished glucose intake or excessive utilization of glucose during exercise and other stressful situations, the glucagon mechanism is very valuable.

3. Also, in hypoglycemia, a direct effect of low blood glucose on the hypothalamus stimulates the sympathetic nervous system. In turn, the epinephrine secreted by the adrenal glands causes still further release of glucose from the liver This, too, helps protect against severe hypoglycemia

4. And, finally, over a period of hours and days, both growth hormone and cortisol are secreted in response to prolonged hypoglycemia, and they both decrease the rate of glucose utilization by most cells of the body. This, too, helps return the blood glucose concentration toward normal.

Importance of Blood Glucose Regulation. One might ask the question· Why is it so important to maintain a constant blood glucose concentration, particularly since most tissues can shift to utilization of fats and proteins for energy in the absence of glucose? The answer is that glucose is the only nutrient that normally can be utilized by the *brain, retina,* and *germinal epithelium of the gonads* in sufficient quantities to supply them with their required energy. Therefore, it is important to maintain the blood glucose concentration at a sufficiently high level to provide this necessary nutrition.

Most of the glucose formed by gluconeogenesis *during the interdigestive period is used for metabolism* in the brain. Indeed, it is important that the pancreas not secrete any insulin during this time, for otherwise the scant supplies of glucose that are available would all go into the muscles and other peripheral tissues, leaving the brain without a nutritive source.

On the other hand, it is also important that the blood glucose concentration not rise too high for three reasons: First, glucose exerts a large amount of osmotic pressure in the extracellular fluid, and, if the glucose concentration rises to excessive values, this can cause considerable cellular dehydration. Second, an excessively high level of blood glucose concentration causes loss of glucose in the urine. And third, this causes osmotic diuresis by the kidneys, which can deplete the body of its fluids and electrolytes.

DIABETES MELLITUS

Etiology. Diabetes mellitus is caused in almost all instances by diminished rates of secretion of insulin by the beta cells of the islets of Langerhans. It is usually divided into two different types: *juvenile diabetes* that usually, but not always, begins in early life, and *maturity-onset diabetes* that usually, but not always, begins in later life and mainly in obese persons.

Heredity plays an important role in the development of both these types of diabetes. The juvenile type is usually rapid in onset and seems to result from hereditary predisposition to (a) development of antibodies against the beta cells, thus causing autoimmune destruction of these, (b) possible destruction of the beta cells by viral disease, or (c) possible simple degeneration of these cells.

The maturity-onset type of diabetes seems to result from degeneration or suppression of the beta cells as a result of more rapid aging in susceptible persons than in others. Obesity predisposes to this type of diabetes, probably for two different reasons: (1) Though the cause is not understood, in obesity the beta cells of the islets of Langerhans become less responsive to stimulation by increased blood glucose. Therefore, the surge of insulin secretion following a meal is less marked in obese persons. (2) Obesity also greatly decreases the number of insulin receptors in the insulin target cells throughout the body For these reasons, increased quantities of insulin are required to have the same metabolic effects in obese persons as in nonobese persons. Therefore, in most persons who have the obesity type of diabetes, the condition can often be treated simply by dietary control of the obesity itself.

PATHOLOGICAL PHYSIOLOGY OF DIABETES MELLITUS

Most of the pathology of diabetes mellitus can be attributed to one of the following three major effects of insulin lack. (1) decreased utilization of glucose by the body cells, with a resultant increase in blood glucose concentration to as high as 300 to 1200 mg/dl; (2) markedly increased mobilization of fats from the fat storage areas, causing abnormal fat metabolism as well as deposition of lipids in vascular walls to cause atherosclerosis; and (3) depletion of protein in the tissues of the body.

However, in addition, some special pathological physiological problems occur in diabetes mellitus that are not so readily apparent These are

Loss of Glucose in the Urine of the Diabetic Person. Whenever the quantity of glucose entering the kidney tubules in the glomerular filtrate rises above approximately 225 mg per minute, a significant proportion of the excess glucose then cannot be reabsorbed and instead spills into the urine. This normally occurs when the blood glucose level rises over 180 mg/dl, which is called the blood "threshold" for the appearance of glucose in the urine. When the blood glucose level rises to 300 to 500 mg/dl—common values in persons with severe untreated diabetes—a hundred or more grams of glucose can be lost into the urine each day.

Dehydrating Effect of Elevated Blood Glucose Levels in Diabetes. Blood glucose levels as high as 1200 mg/dl, 12 times normal, can occur under certain conditions in extreme diabetes. Yet the only significant effect of the elevated glucose is dehydration of the tissue cells, for glucose does not diffuse easily through the pores of the cell membrane, and the increased osmotic pressure in the extracellular fluids ca
fer of water out of the cells.

In addition to the direct dehydrating effect of excessive glucose, the loss of glucose in the urine causes *osmotic diuresis* because the osmotic effect of glucose in the tubules prevents tubular reabsorption of fluid. The overall effect is dehydration of the extracellular fluid, which then causes compensatory dehydration of the intracellular fluid for reasons discussed in Chapter 33. Thus, one of the important features of diabetes is a tendency for both extracellular and intracellular dehydration to develop, and these are also often associated with *circulatory shock*.

Acidosis in Diabetes. The shift from carbohydrate to fat metabolism in diabetes has already been discussed. When the body depends almost entirely on fat for energy, the level of acetoacetic acid and β-hydroxybutyric acid in the body fluids may rise from 1 mEq/liter to as high as 10 mEq/liter. This, obviously, is likely to result in acidosis.

A second effect, which is usually even more important in causing acidosis than is the direct increase in keto acids, is a decrease in sodium concentration caused by the following effect: Keto acids have a low threshold for excretion by the kidneys, therefore, when the keto acid level rises in diabetes, as much as 100 to 200 grams of keto acids can be excreted in the urine each day. Because these are strong acids, having a pK averaging 4.0 or less, very little of them can be excreted in the acidic form but instead is excreted combined with sodium derived from the extracellular fluid. As a result, the sodium concentration in the extracellular fluid usually decreases, and the sodium is replaced by increased quantities of hydrogen ions, thus adding greatly to the acidosis.

Obviously, all the usual reactions that occur in metabolic acidosis take place in diabetic acidosis. These include *rapid and deep breathing* called "Kussmaul respiration," which causes excessive expiration of carbon dioxide, and *marked decrease in bicarbonate content of the extracellular fluids*. Likewise, *large quantities of chloride ion are excreted by the kidneys* as an additional compensatory mechanism for correction of the acidosis. Though these extreme effects occur only in the most severe instances of uncontrolled diabetes, they can lead to acidotic coma and death within hours when they do occur. The overall changes in the electrolytes of the blood as a result of severe diabetic acidosis are illustrated in Figure 78–10.

Relationship of Other Diabetic Symptoms to the Pathological Physiology of Insulin Lack. *Polyuria* (excessive elimination of urine), *polydipsia* (excessive drinking of water), *polyphagia* (excessive eating), *loss of weight*, and *asthenia* (lack of energy) are the earliest symptoms of diabetes. As explained, the polyuria is due to the osmotic diuretic effect of glucose in the kidney tubules. In turn, the polydipsia is due to dehydration resulting from polyuria. The failure of glucose (and protein) utilization by the body causes loss of weight and a tendency toward polyphagia. The asthenia apparently also is caused mainly by loss of body protein.

PHYSIOLOGY OF DIAGNOSIS

The usual methods for diagnosing diabetes are based on various chemical tests of the urine and the blood.

Urinary Sugar. Simple office tests or more complicated quantitative laboratory tests may be used for [illegible]ing the quantity of glucose lost in the urine. In

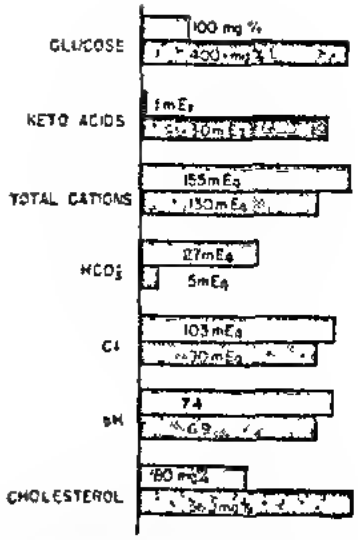


Figure 78–10. Changes in blood constituents in diabetic coma showing normal values (light bars) and diabetic values (dark bars).

general, the normal person loses undetectable amounts of glucose, whereas the diabetic loses glucose in small to large amounts, in proportion to the severity of disease and the intake of carbohydrates. (However, a condition known as *renal glycosuria* sometimes occurs even in persons without diabetes mellitus. This condition results from a low tubular maximum for glucose, as explained in Chapter 38, so that even though the blood glucose level is perfectly normal, a large quantity of glucose may still be lost in the urine.)

The Fasting Blood Glucose Level. The fasting blood sugar level in the early morning, at least eight hours after any previous meal, is normally 80 to 90 mg/dl, and 110 mg/dl is generally considered to be the upper limit of normal. A fasting blood sugar level above this value often indicates diabetes mellitus or, less commonly, either pituitary diabetes or adrenal diabetes.

The Glucose Tolerance Test. As illustrated by the bottom curve in Figure 78–11, when a normal, fasting person ingests 1 gram of glucose per kilogram of body weight, the blood glucose level rises from approximately 90 mg/dl to 120 to 140 mg/dl and falls back to below normal within about two hours.

Though an occasional diabetic person has a normal fasting blood glucose concentration, it is usually above 110 mg/dl and the glucose tolerance test is almost always abnormal. On ingestion of glucose, these persons exhibit a much greater than normal rise in blood glucose level, as illustrated by the upper curve in Figure 78–11; the glucose level falls back to the control value only after some four to six hours, and it fails to fall below the control level. This slow fall of the curve and its failure to fall below the control level illustrates that the normal increase in insulin secretion following glucose ingestion does not occur in the diabetic person, and a diagnosis of diabetes mellitus can usually be definitely established on the basis of such a curve.

Acetone Breath. As pointed out in Chapter 68, small quantities of acetoacetic acid, which increase greatly in severe diabetes, can be converted to acetone, which is volatile and is vaporized into the expired air. Consequently, one frequently can make a diagnosis of diabetes mellitus simply by smelling acetone on the breath of a patient. Also, keto acids can be detected by chemical

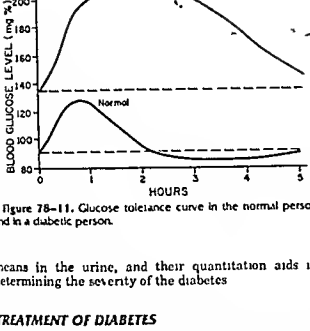


Figure 78–11. Glucose tolerance curve in the normal person and in a diabetic person.

means in the urine, and their quantitation aids in determining the severity of the diabetes

TREATMENT OF DIABETES

The theory of treatment of diabetes mellitus is to administer enough insulin so that the patient will have as nearly normal carbohydrate, fat, and protein metabolism as possible. Optimal therapy can prevent most acute effects of diabetes and greatly delay the chronic effects as well.

Insulin is available in several different forms "Regular" insulin has a duration of action lasting from 3 to 8 hours, whereas other forms of insulin (precipitated with zinc or with various protein derivatives) are absorbed slowly from the injectate site and therefore have effects that last as long as 10 to 48 hours Ordinarily, the severely diabetic patient is given a single dose of one of the longer-acting insulins each day; this increases overall carbohydrate metabolism throughout the day Then additional quantities of regular insulin are given at those times of the day when the blood glucose level tends to rise too high, such as at meal times Thus, each patient is established on an individualized pattern of treatment.

Diet of the Diabetic. The insulin requirements of a diabetic are established with the patient on a standard diet containing normal, well-controlled amounts of carbohydrates, and any change in the quantity of carbohydrate intake changes the requirements for insulin In the normal person, the pancreas has the ability to adjust the quantity of insulin produced to the intake of carbohydrate; but in the completely diabetic person, this control function is totally lost.

In the obesity maturity-onset type of diabetes, the disease can often be controlled by weight reduction alone. The decreased fat reduces the insulin requirements, and the pancreas can now supply the need.

Control of the Diabetic Patient in Fever and Exercise. Any abnormal state of metabolism in the diabetic patient alters the insulin requirement Thus, fever, severe infection, and so forth frequently increase the requirement for insulin immensely, and failure to give the extra insulin can cause diabetic coma.

On the other hand, exercise frequently has exactly the opposite effect on insulin requirement, often reducing the requirement below that usually needed. The peculiar reason for this is that increased muscle activity increases the transport of glucose into the muscle cells even in the absence of insulin. Thus, *exercise actually has an insulin-like effect*. It is often remarked that children with severe diabetes require less insulin to control it if they live a very active life than if they lead an overprotected life.

Relationship of Treatment to Arteriosclerosis. Diabetic patients, mainly because of their high levels of circulating cholesterol and other lipids, develop atherosclerosis, orteriosclerosis, severe coronary heart disease, and multiple microcirculatory lesions far more easily than do normal persons. Indeed, those who have relatively poorly controlled diabetes throughout childhood are likely to die of heart disease in their twenties.

In the early days of treating diabetes it was the tendency to reduce severely the carbohydrates in the diet so that the insulin requirements would be minimized This procedure kept the blood sugar level down to normal values and prevented loss of glucose in the urine, but it did not prevent the abnormalities of fat metabolism. Consequently, there is a tendency at present to allow the patient an almost normal carbohydrate diet and then to give simultaneously large quantities of insulin to metabolize the carbohydrates This depresses the rate of fat metabolism and also depresses the high level of blood cholesterol

Because the complications of diabetes—such as atherosclerosis, greatly increased susceptibility to infection, diabetic retinopathy, cataracts, hypertension, and chronic renal disease—are more closely associated with the level of the blood lipids than with the level of blood glucose, it is the object of many clinics treating diabetes to administer sufficient glucose and insulin so that the quantity of blood lipids becomes normal.

DIABETIC COMA

If diabetes is not controlled satisfactorily, severe dehydration and acidosis may result, and sometimes, even when the person is receiving treatment, sporadic changes in metabolic rates of the cells, such as might occur during bouts of fever, can also precipitate dehydration and acidosis

If the pH of the body fluids falls below approximately 7.0, the diabetic person is likely to develop coma Also, in addition to the acidosis, dehydration is believed to exacerbate the coma Once the diabetic person reaches this stage, the outcome is usually fatal unless immediate treatment is provided.

Physiological Basis of Treating Diabetic Coma. The patient with diabetic coma is extremely refractory to insulin because acidic plasma has an *insulin antagonist*, an alpha globulin, that opposes the action of the insulin. Also, the very high free fatty acid and acetoacetic acid levels in the blood inhibit cellular usage of glucose, as was discussed earlier. Therefore, instead of the usual 60 to 80 units of insulin per day, which is the dosage usually necessary for control of severe diabetes, several times this much insulin must often be given the first day of treatment of coma

Administration of insulin alone is not likely to be sufficient to reverse the abnormal physiology and to effect a cure. In addition, it is usually necessary to correct both the dehydration and acidosis immediately. The dehydration is ordinarily corrected rapidly by administering large quantities of sodium chloride solution,

and the acidosis is often corrected by administering sodium bicarbonate or sodium lactate solution.

HYPERINSULINISM

Though much rarer than diabetes, increased insulin production, which is known as *hyperinsulinism*, does occasionally occur. This usually results from an adenoma of an islet of Langerhans. About 10 to 15 per cent of these adenomas are malignant, and occasionally metastases from the islets of Langerhans spread throughout the body, causing tremendous production of insulin by both the primary and the metastatic cancers. Indeed, in order to prevent hypoglycemia, in some of these patients more than 1000 grams of glucose have had to be administered each 24 hours.

Diagnosis of hyperinsulinism is made with assurance by measuring very high levels of plasma insulin using the radioimmunoassay procedure—especially when the insulin remains high constantly throughout the day and without rising significantly with increased carbohydrate intake.

Insulin Shock and Hypoglycemia. As already emphasized, the central nervous system derives essentially all its energy from glucose metabolism, and insulin is not necessary for this use of glucose. However, if insulin causes the level of blood glucose to fall to low values, the metabolism of the central nervous system becomes depressed. Consequently, in patients with hyperinsulinism, or in diabetic patients who administer too much insulin to themselves, the syndrome called "insulin shock" may occur as follows.

As the blood sugar level falls into the range of 50 to 70 mg/dl, the central nervous system usually becomes quite excitable, for this degree of hypoglycemia seems to facilitate neuronal activity. Sometimes various forms of hallucinations result, but more often the patient simply experiences extreme nervousness, trembles all over, and breaks out in a sweat. As the blood glucose level falls to 20 to 50 mg/dl, clonic convulsions and loss of consciousness are likely to occur. As the glucose level falls still lower, the convulsions cease, and only a state of coma remains. Indeed, at times it is difficult to distinguish between diabetic coma as a result of insulin lack and coma due to hypoglycemia caused by excess insulin. However, the acetone breath and the rapid, deep breathing of diabetic coma are not present in hypoglycemic coma.

Obviously, proper treatment for a patient who has hypoglycemic shock or coma is immediate intravenous administration of large quantities of glucose. This usually brings the patient out of shock within a minute or more. Also, administration of glucagon (or, less effectively, epinephrine) can cause glycogenolysis in the liver and thereby increase the blood glucose level extremely rapidly.

If treatment is not effected immediately, permanent damage to the neuronal cells of the central nervous system occurs, this happens especially in prolonged hyperinsulinism due to pancreatic tumors.

REFERENCES

Bennett, P. H. The diagnosis of diabetes. New international classification and diagnostic criteria. *Annu. Rev. Med.*, *34*:295, 1983.

Cudworth, A. G., *et al.* Etiology of Type I diabetes mellitus. Heterogeneity and immunological events leading to clinical onset. *Annu. Rev. Med.* 34:13, 1983.

Czech, M. P. New perspectives on the mechanism of insulin action. *Recent Prog. Horm. Res.* *40*:317, 1984.

Fain, J. N. Insulin secretion and action. *Metabolism*, *33*:672, 1984.

Fein, F. S., and Sonnenblick, E. H. Diabetic cardiomyopathy. *Prog. Cardiovasc. Dis.* *27*:255, 1985.

Fitzgerald, P. J., and Morrison, A. B. (eds.) The Pancreas. Baltimore, Williams & Wilkins, 1980.

Flier, J. S. Insulin receptors and insulin resistance. *Annu. Rev. Med.* *34*:145, 1983.

Gammeltoft, S. Insulin receptors. Binding kinetics and structure-function relationship of insulin. *Physiol. Rev.* 64:1321, 1984.

Guyton, J. R., *et al.* A model of glucose-insulin homeostasis in man that incorporates the heterogeneous fast pool theory of pancreatic insulin release. *Diabetes*, *27*:1027, 1978.

Hedeskov, C. J. Mechanism of glucose induced insulin secretion. *Physiol. Rev.*, *60*:442, 1980.

Hellerstrom, C. The life story of the pancreatic B cell. *Diabetologia* *26*:393, 1984.

Henquin, J. C., and Meissner, H. P. Significance of ionic fluxes and changes in membrane potential for stimulus-secretion coupling in pancreatic B-cells. *Experientia*, *40*:1043, 1984.

Howell, S. L., and Tyhurst, M. Insulin secretion. The effector system. *Experimentia*, *40*:1098, 1984.

Irsigler, K. (ed.) New Approaches to Insulin Therapy. Baltimore, University Park Press, 1981.

Jacobs, S., and Cuatrecasas, P. Insulin receptors. *Annu. Rev. Pharmacol. Toxicol.* *23*:461, 1983.

Klachko, D. M., *et al.* (eds.) The Endocrine Pancreas and Juvenile Diabetes. New York, Plenum Press, 1979.

Kolb, H. (ed.) Diabetes and Immunology; Pathogenesis and Immunotherapy. Minneapolis, J. K. Burgess, 1983.

Kraus-Friedmann, N. Hormonal regulation of hepatic gluconeogenesis. *Physiol. Rev.* *64*:170, 1984.

Krieger, D. T. (ed.) Current Therapy in Endocrinology, 1983–1984. St. Louis, C. V. Mosby, 1983.

Lenzen, S., and Bailey, C. J. Thyroid hormones, gonadal and adrenocortical steroids and the function of the islets of Langerhans. *Endocr. Rev.* *5*:413, 1984.

Malaisse, W. J., *et al.* Coupling factors in nutrient-induced insulin release. *Experientia* *40*:1035, 1984.

Notkins, A. L. The cause of diabetes. *Sci. Am.*, *241*(5):62, 1979.

Pace, C. S. Role of pH as a transduction device in triggering electrical and secretory responses in islet B cells. *Fed. Proc.*, *43*:2379, 1984.

Park, C. R., *et al.* The action of insulin on the transport of glucose through the cell membrane. *Am. J. Med.*, *26*:674, 1959.

Park, C. R., *et al.* The regulation of glucose uptake in muscle as studied in the perfused rat heart. *Recent Prog. Horm. Res.*, *17*:493, 1961.

Permutt, M. A., et al. Insulin gene structure and function. A review of studies using recombinant DNA methodology. *Diabetes Care*, *7*:386, 1984.

Pipeleers, D. Islet cell interactions with pancreatic B-cells. *Experientia* *40*:1114, 1984.

Post, R. L., *et al.* Regulation of glucose uptake in muscle. III. The interaction of membrane transport and phosphorylation in the control of glucose uptake. *J. Biol. Chem.* *236*:269, 1961.

Robertson, R. P. Prostaglandins, glucose homeostasis, and diabetes mellitus. *Annu. Rev. Med.*, *34*:1, 1983.

Sutherland, D. E. R. Transplantation of the pancreas. Current status. *Annu. Rev. Med.*, *33*:271, 1982.

Unger, R. H., and Orci, L. Glucagon secretion, transport, metabolism, physiologic regulation of secretion, and derangements in diabetes. *In* DeGroot, L. J., *et al.* (eds.) Endocrinology, Vol. 2. New York, Grune & Stratton, 1979, p. 959.

Ward, W. K., *et al.* Pathophysiology of insulin secretion in noninsulin-dependent diabetes mellitus. *Diabetes Care*, *7*:491, 1984.

Wollheim, C. B., and Sharp, G. W. G. Regulation of insulin release by calcium. *Physiol. Rev.*, *61*:914, 1981.

79

Parathyroid Hormone, Calcitonin, Calcium and Phosphate Metabolism, Vitamin D, Bone, and Teeth

The physiology of parathyroid hormone and of the hormone calcitonin is closely related to calcium and phosphate metabolism, the function of vitamin D, and the formation of bone and teeth. Therefore, these are discussed together in the present chapter.

CALCIUM AND PHOSPHATE IN THE EXTRACELLULAR FLUID AND PLASMA—FUNCTION OF VITAMIN D

ABSORPTION AND EXCRETION OF CALCIUM AND PHOSPHATE

Intestinal Absorption of Calcium and Phosphate. Calcium is poorly absorbed form the intestinal tract because of the relative insolubility of many of its compounds and also because bivalent cations are poorly absorbed through the intestinal mucosa anyway. On the other hand, phosphate is absorbed exceedingly well most of the time except when excess calcium is in the diet, the calcium tends to form almost insoluble calcium phosphate compounds that fail to be absorbed but instead pass on through the bowels to be excreted in the feces.

Excretion of Calcium in Feces and Urine; Net Rate of Absorption. Above five sixths of the daily intake of calcium is excreted in the feces, and the remaining sixth in the urine. The approximate daily turnover rates for calcium in the adult are the following:

Intake	800 mg
Intestinal absorption	350 mg.
Secretion in gastrointestinal juices	190 mg.
Net absorption over secretion	170 mg.
Loss in the feces	630 mg.
Excretion in the urine	170 mg.

Excretion of calcium in the urine conforms to much the same principles as sodium excretion. All but a few per cent of the calcium in the glomerular filtrate is reabsorbed in the proximal tubules and ascending limbs of the loops of Henle. Then in the distal tubules and collecting ducts, further reabsorption of the remaining calcium is very selective, depending upon the calcium ion concentration in the blood When low, this reabsorption is very great so that almost no calcium is lost in the urine On the other hand, even a minute increase in calcium ion concentration above normal increases calcium excretion markedly We shall see later in the chapter that one of the most important factors controlling reabsorption of calcium in the distal portions of the nephron, and therefore controlling the rate of calcium excretion, is parathyroid hormone

Intestinal and Urinary Excretion of Phosphate. Except for the portion of phosphate that is excreted in the feces in combination with calcium, almost all the dietary phosphate is absorbed into the blood from the gut and later excreted in the urine

Phosphate is a *threshold substance*, that is, when its concentration in the plasma is below the critical value of approximately 1 millimole/liter, no phosphate at all is lost into the urine; but, above this critical concentration, the rate of phosphate loss is directly proportional to the additional increase. Thus, the kidney regulates the phosphate concentration in the extracellular fluid by altering the rate of phosphate excretion in accordance with the plasma phosphate concentration.

However, as discussed later in the chapter, phosphate excretion by the kidneys is greatly increased by parathyroid hormone, thereby playing an important role in the control of plasma phosphate concentration.

VITAMIN D AND ITS ROLE IN CALCIUM AND PHOSPHATE ABSORPTION

Vitamin D has a potent effect on increasing calcium absorption from the intestinal tract; it also has important effects on both bone deposition

and bone reabsorption, as will be discussed later in the chapter. However, vitamin D itself is not the active substance that actually causes these effects. Instead, the vitamin D must first be converted through a succession of reactions in the liver and the kidney to the final active product, *1,25-dihydroxycholecalciferol* (also called 1,25-$(OH)_2$-D3). Figure 79–1 illustrates the succession of steps that leads to the formation of this substance from vitamin D. Let us discuss these steps.

The Vitamin D Compounds. Several different compounds derived from sterols belong to the vitamin D family, and all these perform more or less the same functions. The most important of these, called vitamin D_3, is *cholecalciferol*. Most of this substance is formed in the skin as a result of irradiation of *7-dehydrocholesterol* by ultraviolet rays from the sun. Consequently, appropriate exposure to the sun prevents vitamin D deficiency. The vitamin D compounds in the food are identical to cholecalciferol except for the substitution of one or more atoms that do not affect their function.

Conversion of Cholecalciferol to 25-Hydroxycholecalciferol in the Liver and Its Feedback Control. The first step in the activation of cholecalciferol is to convert it to 25-hydroxycholecalciferol; this occurs in the liver. The process, however, is itself a limited one because the 25-hydroxycholecalciferol has a feedback inhibitory effect on the conversion reactions. This feedback effect is extremely important for two reasons:

First, the feedback mechanism regulates very precisely the concentration of 25-hydroxycholecalciferol in the plasma, an effect that is illustrated

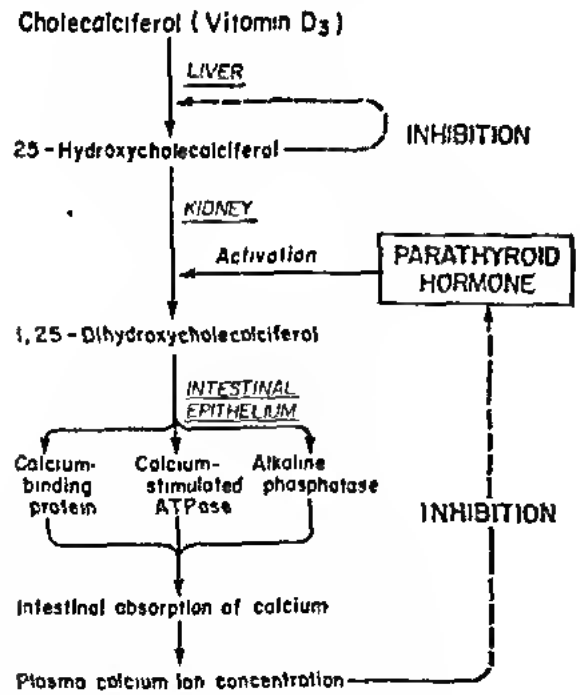


Figure 79–1. Activation of vitamin D, to form *1,25-dihydroxycholecalciferol*, and the role of vitamin D in controlling the plasma calcium concentration.

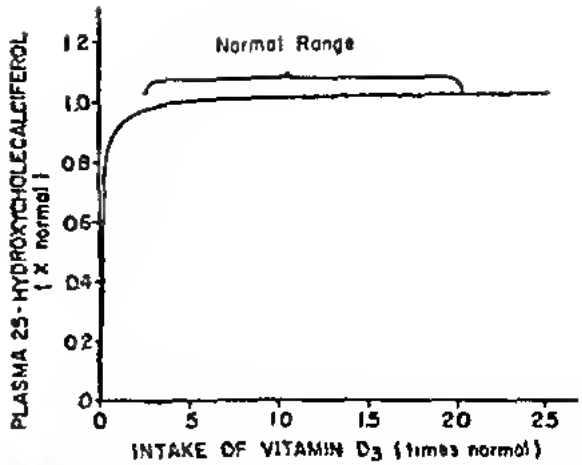


Figure 79–2. Effect of increasing vitamin D_3 intake on the plasma concentration of 25-hydroxycholecalciferol. This figure shows that tremendous changes in vitamin D intake have little effect on the final quantity of activated vitamin D that is formed.

in Figure 79–2. Note that the intake of vitamin D_3 can change manyfold, and yet the concentration of 25-hydroxycholecalciferol still remains within a few per cent of its normal mean value. Obviously this high degree of feedback control prevents excessive action of vitamin D when it is present in too great a quantity.

Second, this controlled conversion of vitamin D, to 25-hydroxycholecalciferol conserves the vitamin D for future use, because once it is converted, it persists in the body for only 2 to 5 days thereafter, whereas in the vitamin D form it can be stored in the liver for as long as several months.

Formation of 1,25-Dihydroxycholecalciferol in the Kidneys and Its Control by Parathyroid Hormone. Figure 79–1 also illustrates the conversion in the kidneys of 25-hydroxycholecalciferol to 1,25-dihydroxycholecalciferol. This latter substance is by far the most active form of vitamin D, for the previous products in the scheme of Figure 79–1 have relatively little vitamin D effect. Therefore, in the absence of the kidneys vitamin D loses much of its effectiveness.

Note also in Figure 79–1 that the conversion of 25-hydroxycholecalciferol to 1,25-dihydroxycholecalciferol requires parathyroid hormone. In the absence of this hormone, either none or almost none of the 1,25-dihydroxycholecalciferol is formed. Therefore, parathyroid hormone exerts a potent effect in determining the functional effects of vitamin D in the body, specifically its effects on calcium absorption in the intestines and its effects on bone.

Effect of Calcium Ion Concentration on Control of the Formation of 1,25-Dihydroxycholecalciferol. Figure 79–3 illustrates that the plasma concentration of 1,25-dihydroxycholecalciferol is inversely affected by the concentration of calcium in the plasma. There are two reasons for this. First, the calcium ion itself has a slight effect in prevention of the conversion of 25-hydroxychole-

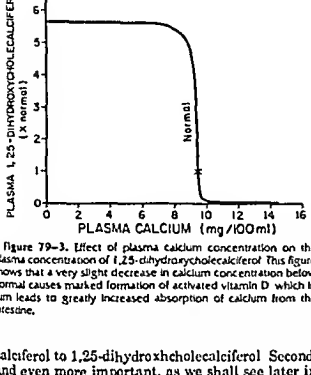


Figure 79–3. Effect of plasma calcium concentration on the plasma concentration of 1,25-dihydroxycholecalciferol. This figure shows that a very slight decrease in calcium concentration below normal causes marked formation of activated vitamin D which in turn leads to greatly increased absorption of calcium from the intestine.

calciferol to 1,25-dihydroxhcholecalciferol. Second, and even more important, as we shall see later in the chapter, the rate of secretion of parathyroid hormone is greatly suppressed when the plasma calcium ion concentration rises above 10 mg/dl. Parathyroid hormone promotes the conversion of 25-hydroxycholecalciferol to 1,25-dihydroxycholecalciferol in the kidney; however, when parathyroid hormone is suppressed, the 25-hydroxycholecalciferol is converted to a slightly different compound—24,25-dihydroxycholecalciferol—that has very little vitamin D effect.

Therefore, when the plasma calcium concentration is already too high, the formation of 1,25-dihydroxycholecalciferol is greatly depressed. Lack of this in turn decreases the absorption of calcium from the intestines, from the bones, and from the renal tubules, thus causing the calcium ion concentration to fall back toward its normal level.

"Hormonal" Effect of 1,25-Dihydroxycholecalciferol on the Intestinal Epithelium in Promoting Calcium Absorption. 1,25-Dihydroxycholecalciferol itself functions as a "hormone" to promote intestinal absorption of calcium. The most important mechanism of this effect is that this hormone causes formation of a *calcium-binding protein* in the intestinal epithelial cells. This protein functions in the brush border of these cells to transport calcium into the cell cytoplasm, and the calcium then moves through the basolateral membrane of the cell, probably by facilitated diffusion. The rate of calcium absorption seems to be directly proportional to the quantity of this calcium-binding protein. Furthermore, this protein remains in the cells for several weeks after the 1,25-dihydroxycholecalciferol has been removed from the body, thus causing a prolonged effect on calcium absorption.

Other effects of this "hormone," 1,25-dihydroxycholecalciferol, that might play a role in promoting calcium absorption are (1) it causes the formation of a calcium-stimulated ATPase in the brush border of the epithelial cells; and (2) it causes the formation of an alkaline phosphatase in the epithelial cells. Unfortunately, the precise details of calcium absorption are still unknown.

Effect of Vitamin D on Phosphate Absorption. Much less is known about the effect of vitamin D on phosphate absorption than on calcium absorption. Also, this is much less important because phosphate is usually absorbed relatively easily anyway. However, phosphate flux through the gastrointestinal epithelium is enhanced by vitamin D. It is believed that this results from a direct effect of 1,25-dihydroxycholecalciferol, but it is possible that it results secondarily from this hormone's action on calcium absorption, the calcium in turn acting as a transport mediator for the phosphate.

THE CALCIUM IN THE PLASMA AND INTERSTITIAL FLUID

The concentration of calcium in the plasma averages approximately 9.4 mg/dl, normally varying between 9.0 and 10.0 mg/dl. This is equivalent to approximately 2.4 millimoles per liter. It is apparent that the calcium level in the plasma is regulated within very narrow limits—and mainly by parathyroid hormone, as discussed later in the chapter.

The calcium in the plasma is present in three different forms, as shown in Figure 79–4. (1) Approximately 40 per cent (1.0 mM/liter) of the calcium is combined with the plasma proteins and consequently is nondiffusible through the capillary membrane. (2) Approximately 10 per cent of the

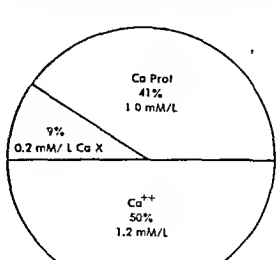


Figure 79–4. Distribution of ionic calcium (Ca^{++}), diffusible but un-ionized calcium (*Ca X*), and calcium proteinate (*Ca Prot*) in blood plasma.

calcium (0.2 mM/liter) is diffusible through the capillary membrane but is combined with other substances of the plasma and interstitial fluids (citrate and phosphate, for instance) in such a manner that it is not ionized. (3) The remaining 50 per cent of the calcium in the plasma is both diffusible through the capillary membrane and ionized. Thus, the plasma and interstitial fluids have a normal *calcium ion concentration of approximately 1 2 mM/liter,* a level only one half the total plasma calcium concentration. This ionic calcium is important for most functions of calcium in the body, including the effect of calcium on the heart, on the nervous system, and on bone formation

THE INORGANIC PHOSPHATE IN THE EXTRACELLULAR FLUIDS

Inorganic phosphate in the plasma is mainly in two forms· HPO_4^{--} AND $H_2PO_4^-$. The concentration of HPO_4^{--} is approximately 1.05 mM/liter and the concentration of $H_2PO_4^-$ approximately 0.26 mM/liter. When the total quantity of phosphate in the extracellular fluid rises, so does the quantity of each of these two types of phosphate ions. Furthermore, when the pH of the extracellular fluid becomes more acid, there is a relative increase in $H_2PO_4^-$ and decrease in the HPO_4^{--} whereas the opposite occurs when the extracellular fluid becomes alkaline. These relationships were presented in the discussion of acid-base balance in Chapter 37.

Because it is difficult to determine chemically the exact quantities of HPO_4^{--} and $H_2PO_4^-$ in the blood, ordinarily the total quantity of phosphate is often expressed in terms of milligrams of *phosphorus* per dl of blood The average total quantity of inorganic phosphorus represented by both phosphate ions is about 4 mg/dl varying between normal limits of 3 to 4 mg/dl in adults and 4 to 5 mg/dl in children

EFFECTS OF ALTERED CALCIUM AND PHOSPHATE CONCENTRATIONS IN THE BODY FLUIDS

Changing the level of phosphate in the extracellular fluid from far below normal to as high as three to four times normal does not cause significant immediate effects on the body.

On the other hand, elevation or depletion of calcium ion in the extracellular fluid causes extreme immediate effects Both chronic hypocalcemia and hypophosphatemia greatly decrease bone mineralization, as explained later in the chapter

Tetany Resulting from Hypocalcemia. When the extracellular fluid concentration of calcium ions falls below normal, the nervous system becomes progressively more and more excitable because this causes increased neuronal membrane permeability to sodium ions The peripheral nerve fibers especially become so excitable that they begin to discharge spontaneously, initiating trains of nerve impulses that pass to the peripheral skeletal muscles and elicit tetanic muscle contraction. Consequently, hypocalcemia causes tetany. But it also occasionally causes convulsions because of its central action of increasing excitability.

Figure 79–5 illustrates tetany in the hand, which usually occurs before tetany develops in most other parts of the body. This is called "carpopedal spasm"

Tetany ordinarily occurs when the blood concentration of calcium falls from its normal level of 9 4 mg to approximately 6 mg/dl which is only 35 per cent below the normal calcium concentration, and it is usually lethal at about 4 mg/dl

In experimental animals, in which the level of calcium can be reduced beyond the normal lethal stage, extreme hypocalcemia can cause marked dilatation of the heart, changes in cellular enzyme activities, increased cell membrane permeability in other cells in addition to nerve cells, and impaired blood clotting.

Hypercalcemia. When the level of calcium in the body fluids rises above normal, the nervous system is depressed, and reflex activities of the central nervous system become sluggish. Also, increased calcium ion concentration decreases the QT interval of the heart, and it causes constipation and lack of appetite, probably because of depressed contractility of the muscle walls of the gastrointestinal tract

The depressive effects of increased calcium level begin to appear when the blood level of calcium rises above approximately 12 mg/dl, and they can become marked as the calcium level rises above 15 mg/dl. When the level of calcium rises above approximately 17 mg/dl in the body fluids, calcium phosphate is likely to precipitate throughout the body, this condition is discussed shortly in connection with parathyroid poisoning

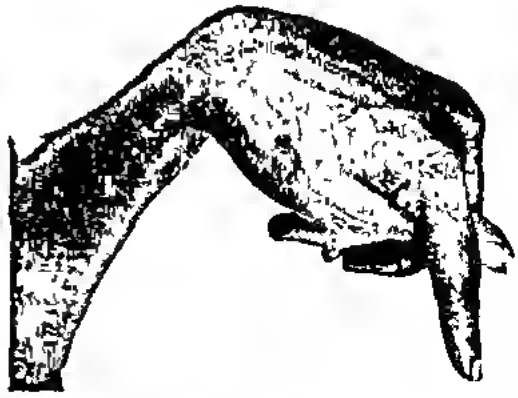


Figure 79–5. Hypocalcemic tetany in the hand, called "carpopedal spasm" (Courtesy of Dr Herbert Langford)

BONE AND ITS RELATIONSHIPS TO EXTRACELLULAR CALCIUM AND PHOSPHATES

Bone is composed of a tough *organic matrix* that is greatly strengthened by deposits of *calcium salts*. Average *compact bone* contains by weight approximately 30 per cent matrix and 70 per cent salts. However, *newly formed bone* may have a considerably higher percentage of matrix in relation to salts.

The Organic Matrix of Bone. The organic matrix of bone is 90 to 95 per cent *collagen fibers*, and the remainder is a homogeneous medium *called ground substance*. The collagen fibers extend primarily along the lines of tensional force. These fibers give bone its powerful tensile strength.

The ground substance is composed of extracellular fluid plus *proteoglycans*, especially *chondroitin sulfate* and *hyaluronic acid*. The precise function of these is not known, though perhaps they help to control the deposition of calcium salts.

The Bone Salts. The crystalline salts deposited in the organic matrix of bone are composed principally of *calcium* and *phosphate*, and the formula for the major crystalline salts, known as *hydroxyapatites*, is the following.

$$Ca^{\cdot\cdot}_{10-x}(H_3O^{+})_{2x} \cdot (PO_4)_6(OH^{-})_2$$

Each crystal—about 400 Å long, 10 to 30 Å thick, and 100 Å wide—is shaped like a long, flat plate. The relative ratio of calcium to phosphorus can vary markedly under different nutritional conditions, the Ca/P ratio on a weight basis varying between 1.3 and 2.0.

Magnesium, sodium, potassium, and *carbonate* ions are also present among the bone salts, though x-ray diffraction studies fail to show definite crystals formed by these. Therefore, they are believed to be conjugated to the hydroxyapatite crystals rather than organized into distinct crystals of their own. This ability of many different types of ions to conjugate to bone crystals extends to many ions normally foreign to bone, such as *strontium, uranium, plutonium, the other transuranic elements, lead, gold, other heavy metals*, and *at least 9 of 14 of the major radioactive products released by explosion of the hydrogen bomb*. Deposition of radioactive substances in the bone can cause prolonged irradiation of the bone tissues, and, if a sufficient amount is deposited, an osteogenic sarroma (bone cancer) almost invariably eventually develops.

Tensile and Compressional Strength of Bone. Each collagen fiber of *compact* bone is composed of repeating periodic segments every 640 A along its length; hydroxyapatite crystals lie adjacent to each segment of the fiber, bound tightly to it. This intimate bonding prevents "shear" in the bone; that is, it prevents the crystals and collagen fibers from slipping out of place, which is essential in providing strength to the bone. In addition, the segments of adjacent collagen fibers overlap each other, also causing hydroxyapatite crystals to be overlapped like bricks keyed to each other in a brick wall.

The collagen fibers of bone, like those of tendons, have great tensile strength, while the calcium salts, which are similar in physical properties to marble, have great compressional strength. These combined properties, plus the degree of bondage between the collagen fibers and the crystals, provide a bony structure that has both extreme tensile and compressional strength. Thus, bones are constructed in exactly the same way that reinforced concrete is constructed. The steel of reinforced concrete provides the tensile strength, while the cement, sand, and rock provide the compressional strength. Indeed, the compressional strength of bone is greater than that of even the best reinforced concrete, and the tensile strength approaches that of reinforced concrete.

PRECIPITATION AND ABSORPTION OF CALCIUM AND PHOSPHATE IN BONE—EQUILIBRIUM WITH THE EXTRACELLULAR FLUIDS

Supersaturated State of Calcium and Phosphate Ions in the Extracellular Fluids with Respect to Hydroxyapatite. The concentrations of calcium and phosphate ions in extracellular fluid are considerably greater than those required to cause precipitation of hydroxyapatite. However, inhibitors are present in most tissues of the body, as well as in plasma, to prevent such precipitation; one such inhibitor is pyrophosphate. Therefore, hydroxyapatite crystals fail to precipitate in normal tissues except in bone despite the state of supersaturation of the ions.

Mechanism of Bone Calcification. The initial stage in bone production is the secretion of collagen molecules (called collagen monomers) and ground substance by *osteoblasts*. The collagen monomers polymerize rapidly to form collagen fibers, and the resultant tissue becomes *osteoid*, a cartilage-like material but differing from cartilage in that calcium salts precipitate in it. As the osteoid is formed, some of the osteoblasts become entrapped in the osteoid and then are called *osteocytes*.

Within a few days after the osteoid is formed, calcium salts begin to precipitate on the surfaces of the collagen fibers. The precipitates appear at periodic intervals along each collagen fiber, forming minute nidi that rapidly multiply and grow over a period of days and weeks into the finished product, *hydroxyapatite crystals*.

The initial calcium salts to be deposited probably are not hydroxyapatite crystals but, instead, amorphous compounds (noncrystalline), a probable mixture of such salts as $CaHPO_4 \cdot 2H_2O$, $Ca_3(PO_4)_2 \cdot 3H_2O$, and others. Then by a process of

substitution and addition of atoms, or reabsorption and reprecipitation, these salts are converted into the hydroxyapatite crystals over a period of weeks or months. Yet, as much as 20 to 30 per cent may remain permanently in the amorphous form. This is important, because these salts can be absorbed rapidly when there is need for extra calcium in the extracellular fluid

It is still not known what causes calcium salts to be deposited in osteoid. One theory holds that at the time of formation the collagen fibers are specially constituted in advance for causing precipitation of calcium salts. One variant of this theory suggests that the osteoblasts secrete a substance into the osteoid to neutralize an inhibitor (perhaps pyrophosphate) that normally prevents hydroxyapatite crystallization. Once the pyrophosphate has been neutralized, then the natural affinity of the collagen fibers for calcium salts supposedly causes the precipitation. In support of this theory is the fact that properly prepared collagen fibers from other tissues of the body besides bone will also cause precipitation of hydroxyapatite crystals from plasma.

Precipitation of Calcium in Nonosseous Tissues Under Abnormal Conditions. Though calcium salts almost never precipitate in normal tissues besides bone, under abnormal conditions they do precipitate. For instance, they precipitate in arterial walls in the condition called arteriosclerosis and cause the arteries to become bonelike tubes. Likewise, calcium salts frequently deposit in degenerating tissues or in old blood clots. Presumably, in these instances, the inhibitor factors that normally prevent deposition of calcium salts disappear from the tissues, thereby allowing precipitation.

EXCHANGEABLE CALCIUM

If soluble calcium salts are injected intravenously, the calcium ion concentration can be made to increase immediately to very high levels. However, within a half hour to an hour or more, the calcium ion concentration returns to normal. Likewise, if large quantities of calcium ions are removed from the circulating body fluids, the calcium ion concentration again returns to normal within a half hour to an hour or so. These effects result partly from the fact that the body contains a type of *exchangeable* calcium that is always in equilibrium with the calcium ions in the extracellular fluids. A small portion of this exchangeable calcium is that calcium found in all tissue cells, especially in highly permeable types of cells such as those of the liver and the gastrointestinal tract. However, most of the exchangeable calcium, as shown by studies using radioactively tagged calcium, is in the bone, and it normally amounts to about 0.4 to 1.0 per cent of the total bone calcium. Most of this calcium is probably deposited in the bones in the form of readily mobilizable salts such as $CaHPO_4$ and the other amorphous salts.

The importance of exchangeable calcium to the body is that it provides a rapid buffering mechanism to keep the calcium ion concentration in the extracellular fluids from rising to excessive levels or falling to very low levels under transient conditions of excess or hypoavailability of calcium.

DEPOSITION AND ABSORPTION OF BONE—REMODELING OF BONE

Deposition of Bone by the Osteoblasts. Bone is continually being deposited by *osteoblasts*, and it is continually being absorbed where *osteoclasts* are active. Osteoblasts are found on the outer surfaces of the bones and in the bone cavities A small amount of osteoblastic activity occurs continually in all living bones (on about 4 per cent of all surfaces at any given time, as shown in Figure 79-6) so that at least some new bone is being formed constantly

Absorption of Bone—Function of the Osteoclasts. Bone is also being continually absorbed in the presence of osteoclasts, large multinucleated cells that are normally active at any one time on less than 1 per cent of the bone surfaces Later in the chapter we will see that parathyroid hormone controls the bone absorptive activity of osteoclasts

Histologically, bone absorption occurs immediately adjacent to the osteoclasts, as illustrated in Figure 79-6 The mechanism of this absorption is believed to be the following The osteoclasts send out villus-like projections toward the bone and from these "villi" secrete two types of substances: (1) proteolytic enzymes, released from the lysosomes of the osteoclasts, and (2) several acids, including citric acid and lactic acid. The enzymes presumably digest or dissolve the organic matrix of the bone, while the acids cause solution of the bone salts.

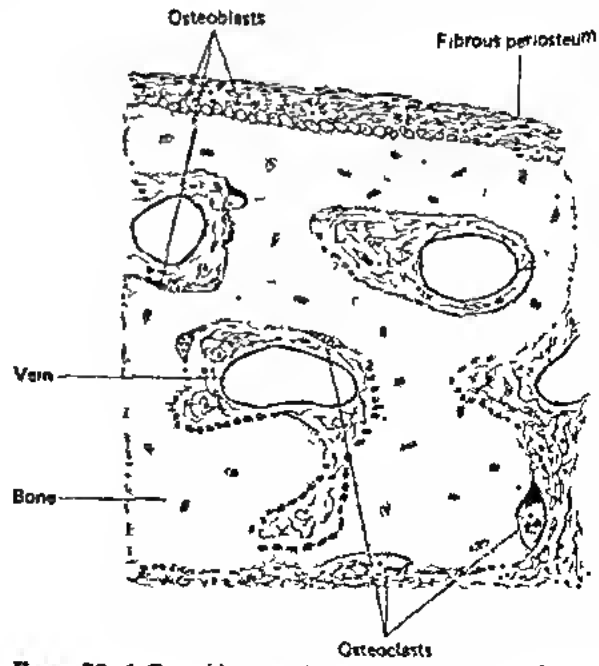


Figure 79-6. Osteoblastic and osteoclastic activity in the same bone

Also, whole fragments of bone salts and collagen are literally gobbled up (phagocytosed) by the "villi" and then digested within the osteoclasts.

Equilibrium Between Bone Deposition and Absorption. Normally, except in growing bones, the rates of bone deposition and absorption are equal to each other so that the total mass of bone remains constant. Usually, osteoclasts exist in small but concentrated masses, and once a mass of osteoclasts begins to develop, it usually eats away at the bone for about three weeks, eating out a tunnel that may be as large as 1 mm in diameter and several millimeters in length. At the end of this time the osteoclasts disappear and the tunnel is invaded by osteoblasts instead; then new bone begins to develop. Bone deposition then continues for several months, the new bone being laid down in successive layers of concentric circles on the inner surfaces of the cavity until the tunnel is filled. Deposition of new bone ceases when the bone begins to encroach on the blood vessels supplying the area. The canal through which these vessels run, called the *haversian canal*, therefore, is all that remains of the original cavity. Each new area of bone deposited in this way is called an *osteon*, as shown in Figure 79–7.

Value of Continual Remodeling of Bone. The continual deposition and absorption of bone has a number of physiologically important functions. First, bone ordinarily adjusts its strength in proportion to the degree of bone stress. Consequently, bones thicken when subjected to heavy loads. Second, even the shape of the bone can be rearranged for proper support of mechanical forces by deposition and absorption of bone in accordance with stress patterns. Third, since old bone becomes relatively weak and brittle, new organic matrix is needed as the old organic matrix degenerates. In this manner the normal toughness of bone is maintained. Indeed, the bones of children, in whom the rates of deposition and absorption are rapid, show little brittleness in comparison with the bones of old age, at which time the rates of deposition and absorption are slow.

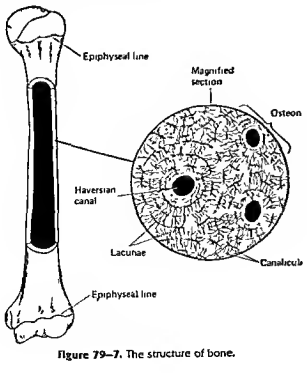


Figure 79–7. The structure of bone.

Control of the Rate of Bone Deposition by Bone "Stress." Bone is deposited in proportion to the compressional load that the bone must carry. For instance, the bones of athletes become considerably heavier than those of nonathletes. Also, if a person has one leg in a cast but continues to walk on the opposite leg, the bone of the leg in the cast becomes thin and as much as 30 per cent decalcified within a few weeks, while the opposite bone remains thick and normally calcified. Therefore, continual physical stress stimulates calcification and osteoblastic deposition of bone.

Bone stress also determines the shape of bones under certain circumstances. For instance, if a long bone of the leg breaks in its center and then heals at an angle, the compression stress on the inside of the angle causes increased deposition of bone, while increased absorption occurs on the outer side of the angle where the bone is not compressed. After many years of increased deposition on the inner side of the angulated bone and absorption on the outer side, the bone can become almost straight, especially in children because of the rapid remodeling of bone at younger ages.

The deposition of bone at points of compressional stress has been suggested to be caused by a *piezoelectric* effect, as follows: Compression of bone causes a negative potential in the compressed areas and a positive potential elsewhere in the bone. It has been shown that minute quantities of current flowing in bone cause osteoblastic activity at the negative end of the current flow, which could explain the increased bone deposition at compression sites. On the other hand, usual osteoclastic activity could account for reabsorption of bone at sites of tension.

Repair of a Fracture. A fracture of a bone in some way maximally activates all the periosteal and intraosseous osteoblasts involved in the break. Also, immense numbers of new osteoblasts are formed almost immediately from so-called *osteoprogenitor cells*, which are bone stem cells. Therefore, within a short time a large bulge of osteoblastic tissue and new organic bone matrix, followed shortly by the deposition of calcium salts, develops between the two broken ends of the bone. This is called a *callus*.

Many bone surgeons utilize the phenomenon of bone stress to accelerate the rate of fracture healing. This is done by use of special mechanical fixation apparatuses for holding the ends of the broken bone together so that the patient can use the bone immediately. This obviously causes stress on the opposed ends of the broken bones, which accelerates osteoblastic activity at the break and often shortens convalescence.

Blood Alkaline Phosphatase as an Indication of the Rate of Bone Deposition. The osteoblasts secrete large quantities of alkaline phosphatase when they are actively depositing bone matrix. This phosphatase is believed either to increase the local concentration of inorganic phosphate or to activate the collagen fibers in such a way that they cause the deposition of calcium salts. Since some alkaline phosphatase diffuses into the blood, the blood level of alkaline phosphatase is usually a good indicator of the rate of bone formation.

PARATHYROID HORMONE

For many years it has been known that increased activity of the parathyroid gland causes rapid absorption of calcium salts from the bones

with resultant hypercalcemia in the extracellular fluid, conversely, hypofunction of the parathyroid glands causes hypocalcemia, often with resultant tetany. Also, parathyroid hormone is important in phosphate metabolism as well as in calcium metabolism

Physiologic Anatomy of the Parathyroid Glands. Normally there are four parathyroid glands in the human being; these are located immediately behind the thyroid gland—one behind each of the upper and each of the lower poles of the thyroid Each parathyroid gland is approximately 6 mm long, 3 mm wide, and 2 mm thick and has a macroscopic appearance of dark brown fat, therefore, the parathyroid glands are difficult to locate during thyroid operations. For this reason, before the importance of these glands was generally recognized, total or subtotal thyroidectomy frequently resulted in total removal of the parathyroid glands as well

Removal of half the parathyroid glands usually causes little physiological abnormality. However, removol of three of four normal glands usually causes transient hypoparathyroidism. But even a small quantity of remaining parathyroid tissue is usually capable of hypertrophying satisfactorily to perform the function of all the glands.

The parathyroid gland of the adult human being, illustrated in Figure 79–8, contains mainly *chief cells* and *oxyphil cells*, but oxyphil cells are absent in many animals and in young human beings. The chief cells secrete most of the parathyroid hormone. The function of the oxyphil cells is not certain; they are perhaps modified chief cells that still secrete some hormone.

Chemistry of Parathyroid Hormone. Parathyroid hormone has been isolated in a pure form. It is first synthesized on the ribosomes in the form of a preprohormone, a polypeptide chain of 115 amino acids This is cleaved first to a prohormone and finally to the hormone itself by the endoplasmic reticulum and Golgi apparatus, and then packaged in secretory granules in the cytoplasm of the cells The final hormone has a molecular weight of approximately 9500 and is composed of 84 amino acids Smaller compounds have also been isolated from the parathyroid glands that exhibit parathyroid hormone activity, but the activity is always

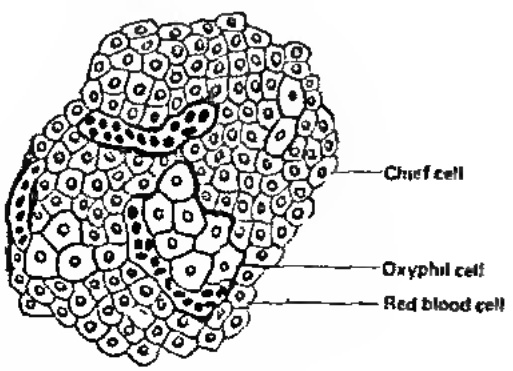


Figure 79–8. Histological structure of a parathyroid gland.

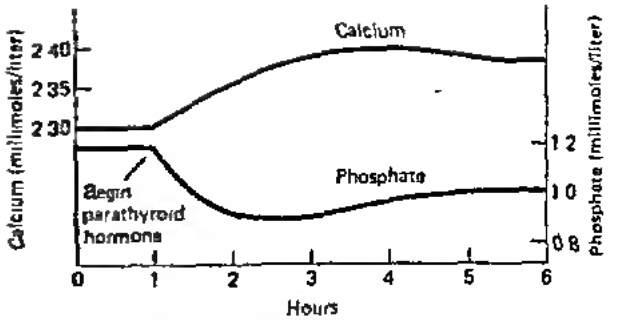


Figure 79–9. Approximate changes in calcium and phosphate concentrations during the first five hours of parathyroid hormone infusion at a moderate rate

slightly to much less than that of the larger protein molecule These are breakdown products of the normal parathyroid hormone

EFFECT OF PARATHYROID HORMONE ON CALCIUM AND PHOSPHATE CONCENTRATIONS IN THE EXTRACELLULAR FLUID

Figure 79–9 illustrates the effect on the blood calcium and phosphate concentrations caused by suddenly beginning to infuse parathyroid hormone into an animal and continuing this for an indefinite period of time Note that at the onset of infusion the calcium ion concentration begins to rise and reaches a plateau level in about 4 hours. On the other hand, the phosphate concentration falls more rapidly and also reaches a depressed plateau level within an hour or two. The rise in calcium concentration is caused principally by two effects (1) an effect of parathyroid hormone in causing calcium and phosphate absorption from the bone, and (2) a rapid effect of parathyroid hormone in decreasing the excretion of calcium by the kidneys. The decline in phosphate concentration, on the other hand, is caused by a very strong effect of parathyroid hormone on the kidney in causing excessive renal phosphate excretion, an effect that is usually great enough to override increased phosphate absorption from the bone.

Calcium and Phosphate Absorption from the Bone Caused by Parathyroid Hormone

Parathyroid hormone seems to have two separate effects on bone in causing absorption of calcium and phosphate. One is a very rapid phase that takes place in minutes and probably results from activation of the already existing bone cells to promote calcium and phosphate absorption. The second phase is a much slower one, requiring several days or even weeks to become fully developed, and it results from proliferation of the osteoclasts, followed by greatly increased osteoclastic reabsorption of the bone itself, not merely absorption of the calcium phosphate salts from the bone

The Rapid Phase of Calcium and Phosphate Absorption—Osteolysis. When large quantities of parathyroid hormone are injected, the calcium ion concentration in the blood begins to rise within minutes, long before any new bone cells can be developed. Histological studies have shown that the parathyroid hormone causes removal of bone salts from the bone matrix in the vicinity of the osteocytes lying within the bone itself and also in the vicinity of the osteoblasts along the bone surface. Yet, strangely enough, one does not usually think of either osteoblasts or osteocytes functioning to cause bone salt absorption, because both these types of cells are osteoblastic in nature and are normally associated with bone deposition and its calcification. However, recent studies have shown that the osteoblasts and osteocytes form a system of interconnected cells that spreads over all the bone surfaces except the small surface areas that are adjacent to the osteoclasts. Also, long filmy processes extend from osteocyte to osteocyte throughout the bone structure, and these processes also connect with the surface osteocytes and osteoblasts. This extensive system is called the *osteocytic membrane system*, and it is believed to provide a membrana that separates the bone itself from the extracellular fluid. Between the osteocytic membrana and the bone is a small amount of fluid called simply *bone fluid*. Indirect experiments indicate that the osteocytic membrane pumps calcium ions from the bone fluid into the extracellular fluid, creating a calcium ion concentration in the bone fluid only one third of that in the extracellular fluid. When the osteocytic pump becomes excessively activated, the bone fluid calcium concentration falls even lower, and calcium phosphate salts are then absorbed from the bone. This effect is called *osteolysis*, and it occurs without absorption of the bone matrix. When the pump is inactivated, the bone fluid calcium concentration rises to a higher level, and calcium phosphate salts are then redeposited in the matrix.

But, where does parathyroid hormone fit into this picture? It seems that parathyroid hormone can activate the calcium pump strongly, thereby causing rapid removal of calcium phosphate salts from the amorphous bone crystals that lie near the bone surfaces and adjacent to the bone fluid. The parathyroid hormone is believed to stimulate this pump by increasing the calcium permeability of the bone fluid side of the osteocytic membrane, thus allowing calcium ions to diffuse into the membrane cells from the bone fluid. Then the calcium pump on the other side of the cell membrane transfers the calcium ions the rest of the way into the extracellular fluid.

The Slow Phase of Bone Absorption and Calcium Phosphate Release—Activation of the Osteoclasts. A much better-known effect of parathyroid hormone, and one for which the evidence is also much clearer, is its ability to activate the osteoclasts. These in turn set about their usual task of gobbling up the bone.

Activation of the osteoclastic system occurs in two stages: (1) immediate activation of the osteoclasts that are already formed, and (2) formation of new osteoclasts from *osteoprogenitor cells*. Usually several days of excess parathyroid hormone cause the osteoclastic system to become well developed, but it can continue to grow for literally months under the influence of very strong parathyroid hormone stimulation.

Parathyroid hormone also transiently depresses osteoblastic activity However, after a few days to a few weeks the osteoclastic resorption of bone leads to weakened bones and secondary stimulation of the osteoblasts Therefore, the late effect is actually to enhance both osteoblastic and osteoclastic activity. Still, even in the late stages, there is more bone absorption than bone deposition.

Bone contains such great amounts of calcium in comparison with the total amount in all the extracellular fluids (about 1000 times as much) that even when parathyroid hormone causes a great rise in calcium concentration in tha fluids, it is impossible to discern any immediate effect on tha bones Yet prolonged administration or secretion of parathyroid hormone—over a period of many months, or even years—finally results in evident absorption in all the bones with development of large cavities filled with very large, multinucleated osteoclasts.

Effect of Parathyroid Hormone on Phosphate and Calcium Excretion by the Kidneys

Administration of parathyroid hormone causes immediate and rapid loss of phosphate in the urine. This effect is caused by diminished proximal tubular reabsorption of phosphate ions.

Parathyroid hormone also increases the reabsorption of calcium at the same time that it diminishes phosphate reabsorption. Moreover, it also increases the rate of reabsorption of magnesium ions and hydrogen ions, while it decreases the reabsorption of sodium, potassium, and amino acid ions in much the same way that it affects phosphate. However, the increased calcium absorption occurs in the *distal tubules* and *collecting ducts* instead of the proximal tubules.

Were it not for the effect of parathyroid hormone on the kidneys to increase calcium reabsorption, continual loss of calcium into the urine would eventually deplete the bones of this mineral.

Effect of Parathyroid Hormone on Intestinal Absorption of Calcium and Phosphate

At this point we should be reminded again that parathyroid hormone greatly enhances both cal-

cium and phosphate absorption from the intestines by increasing the formation of 1,25-dihydroxycholecalciferol from vitamin D, as was discussed earlier in the chapter.

Effect of Vitamin D on Bone and Its Relation to Parathyroid Hormone Activity

Vitamin D plays important roles in both bone absorption and bone deposition. Administration of extreme quantities of vitamin D causes absorption of bone in much the same way that administration of parathyroid hormone does. Also, in the absence of vitamin D, the effect of parathyroid hormone in causing bone absorption is greatly reduced or even prevented. The mechanism of this action of vitamin D is not known, but it is believed to result from the effect of 1,25-dihydroxycholecalciferol (the principal active product of vitamin D) in increasing calcium transport through cellular membranes.

Vitamin D in much smaller quantities promotes bone calcification. Obviously, one of the ways in which it does this is to increase calcium and phosphate absorption from the intestines. However, even in the absence of such increase, it still enhances the mineralization of bone. Here again, the mechanism of the effect is unknown, but it probably also results from the ability of 1,25-dihydroxycholecalciferol to cause transport of calcium ions through cell membranes—but perhaps in the opposite direction through the osteoblastic or osteocytic cell membranes.

Role of Cyclic AMP as a Mediator of Parathyroid Stimulation. A large share of the effect of parathyroid hormone on its target organs is almost certainly mediated by the cyclic AMP *second messenger* mechanism. Within a few minutes after parathyroid hormone administration, the concentration of cyclic AMP increases in the osteoclasts and other target cells. This cyclic AMP, in turn, is probably responsible for such functions as osteoclastic secretion of enzymes and acids to cause bone reabsorption, formation of 1,25-dihydroxycholecalciferol in the kidneys, and so forth. However, likely there are other direct effects of parathyroid hormone that function independently of the second messenger mechanism.

CONTROL OF PARATHYROID SECRETION BY THE CALCIUM ION CONCENTRATION

Even the slightest decrease in calcium ion concentration in the extracellular fluid causes the parathyroid glands to increase their rate of secretion within minutes, and if the decreased calcium concentration persists, the glands will hypertrophy sometimes as much as five-fold or more. For instance, the parathyroid glands become greatly enlarged in *rickets*, in which the level of calcium is usually depressed only a few per cent; also they become greatly enlarged in pregnancy, even though the decrease in calcium ion concentration in the mother's extracellular fluid is hardly measurable; and, they are greatly enlarged during lactation because calcium is used for milk formation.

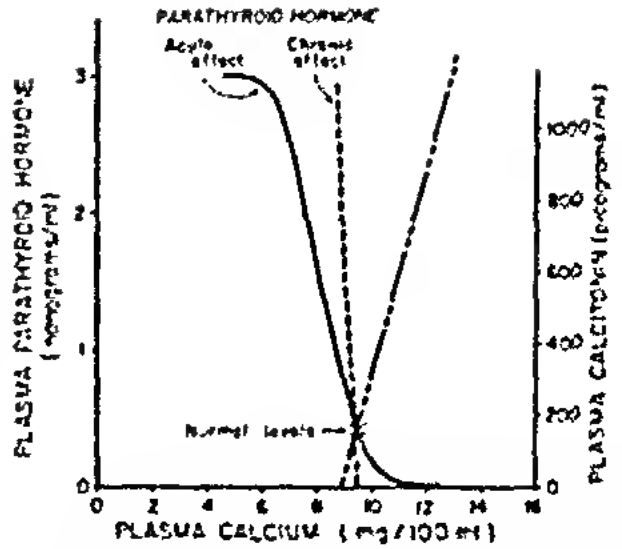


Figure 79-10. Approximate effect of plasma calcium concentration on the plasma concentrations of parathyroid hormone and calcitonin. Note especially that long term, chronic changes in calcium concentration can cause as much as a 100 per cent change in parathyroid hormone concentration for only a few per cent change in calcium concentration.

On the other hand, any condition that increases the calcium ion concentration causes decreased activity and reduced size of the parathyroid glands. Such conditions include (1) excess quantities of calcium in the diet, (2) increased vitamin D in the diet, and (3) bone absorption caused by factors other than parathyroid hormone (for example, bone absorption caused by disuse of the bones).

Figure 79-10 illustrates quantitatively the approximate relationship between plasma calcium concentration and plasma parathyroid hormone concentration. The solid red curve shows the acute relationship when the calcium concentration is changed over a period of a few hours. This shows that a decrease in calcium concentration from 9.4 to 8.4 mg/dl doubles or triples the plasma parathyroid hormone. On the other hand, the chronic relationship that one finds when the calcium ion concentration changes over a period of many weeks, thus allowing time for the glands to hypertrophy, is illustrated by the dashed red line; this illustrates that as little as a 0.1 mg/dl decrease in plasma calcium concentration can double parathyroid hormone secretion. To state this still another way, the chronic relationship between plasma calcium and plasma parathyroid hormone shows that approximately a 1 per cent decrease in calcium can give as much as a 100 per cent increase in parathyroid hormone. Obviously, this is the basis of the body's extremely potent feedback system for control of plasma calcium ion concentration.

CALCITONIN

About 25 years ago, a new hormone that has effects on blood calcium opposite to those of parathyroid hormone was discovered. This hormone,

named *calcitonin* because it reduces the blood calcium ion concentration, is secreted in the human being not by the parathyroid glands but instead by the thyroid gland. In fish, amphibia, reptiles, and birds calcitonin is secreted by the *ultimobranchial glands;* it plays an especially important role in helping control the blood calcium ion concentration when these animals change their habitat from fresh water to sea water, where there are great excesses of calcium. Furthermore, its concentration in these ultimobranchial glands is extremely great. In the human being, ultimobranchial glands do not exist as such but have become incorporated into the thyroid gland. The so-called *parafollicular cells,* or C cells, in the interstitial tissue between the follicles of the human thyroid gland are remnants of the ultimobranchial glands of lower animals, and it is these cells that secrete the calcitonin.

Calcitonin is a large polypeptide with a molecular weight of approximately 3400 and a chain of 32 amino acids.

Effect of Calcitonin in Decreasing Plasma Calcium Concentration. In young animals, calcitonin decreases blood calcium ion concentration very rapidly, beginning within minutes after injection of the calcitonin. Thus, the effect of calcitonin on blood calcium ion concentration is exactly opposite that of parathyroid hormone, and it occurs several times as rapidly.

Calcitonin reduces plasma calcium concentration in three separate ways:

1. The immediate effect is to decrease the osteolytic effect of the osteocytic membrane throughout the bone, thus shifting the balance in favor of deposition of calcium in the rapidly exchangeable pool of bone calcium salts. This effect is especially significant in young animals because of the rapid interchange of absorbed and deposited calcium.

2. The second effect, which can be seen within about an hour, is an increase in osteoblastic activity. However, this is a transient effect, lasting not more than a few days.

3. The third and most prolonged effect of calcitonin is to decrease the formation of new osteoclasts from the osteoprogenitor cells. Also, since osteoclastic resorption of bone leads secondarily to osteoblastic activity, the depressed numbers of osteoclasts are followed by depressed numbers of osteoblasts as well. Therefore, over a long period of time the net result is simply greatly reduced osteoclastic and osteoblastic activity; therefore, there is not a significant prolonged effect on plasma calcium ion concentration. That is, the effect on plasma calcium is mainly a transient one, lasting for a few hours to a few days at most. However, there is a prolonged effect of decreasing the rate of bone remodeling as well as of increasing the amount of calcium salts deposited in bone, under some conditions.

Calcitonin also has minor effects on calcium handling in the kidney tubules and intestinal tract. Again, the effects are opposite those of parathyroid hormone, but quantitatively they appear to be of such little import that they are rarely considered.

Importance of the Calcitonin Effect on Plasma Calcium Concentration. Calcitonin has only a very weak effect on plasma calcium concentration in the adult human being. The reason for this is twofold. First, any initial reduction of the calcium ion concentration caused by calcitonin leads within hours to a powerful stimulation of parathyroid hormone secretion, which almost completely overrides the calcitonin effect. Second, in the adult, the daily rates of absorption and deposition of calcium are very small, and even after the rate of absorption is slowed and deposition increased by calcitonin as much as three to fivefold, this still has only a small effect on plasma calcium ion concentration. On the other hand, the effect in children is much more marked because bone remodeling occurs rapidly in children, with absorption and deposition of calcium as great as 5 or more grams per day—equal to five to ten times the total calcium in all the extracellular fluid. Also, in certain bone diseases such as Paget's disease in which osteoclastic activity is greatly accelerated, calcitonin then has a potent effect of reducing the calcium absorption.

Effect of Plasma Calcium Concentration on the Secretion of Calcitonin

An increase in plasma calcium concentration of about 10 per cent causes an immediate three- to sixfold increase in the rate of secretion of calcitonin, which is illustrated by the dot-dash line of Figure 79–10. This provides a second hormonal feedback mechanism for controlling the plasma calcium ion concentration, but one that works exactly opposite to the parathyroid hormone system.

However, there are two major differences between the calcitonin and the parathyroid feedback systems. First, the calcitonin mechanism operates more rapidly, reaching peak activity in less than an hour, in contrast to the 3 to 4 hours required for peak activity to be attained following the onset of parathyroid secretion.

The second difference is that the calcitonin mechanism acts mainly as a short-term regulator of calcium ion concentration because it is very rapidly overridden by the much more powerful parathyroid control mechanism. Therefore, over a prolonged period of time it is almost entirely the parathyroid system that sets the long-term level of calcium ions in the extracellular fluid. Yet, for short periods, such as for an hour or so after a high calcium meal, calcitonin does seem to play a significant role in decreasing the rise in the calcium ion concentration that otherwise would occur.

When the thyroid gland is removed and calcitonin is no longer secreted, the long-term blood

calcium ion concentration is not measurably altered, but the amount of bone salts may become slightly decreased—again demonstrating the overriding effect of the parathyroid hormonal system.

OVERALL CONTROL OF CALCIUM ION CONCENTRATION

At times the amount of calcium absorbed into or lost from the body fluids is as much as 0.3 gram in an hour. For instance, in cases of diarrhea, several grams of calcium can be secreted in the intestinal juices, passed into the intestinal tract, and lost into the feces each day. Conversely, after ingestion of large quantities of calcium, particularly when there is also an excess of vitamin D activity, a person may absorb as much as 0.3 gram in an hour. This figure compares with a *total quantity of calcium in all the extracellular fluid of about 1 gram.* The addition or subtraction of 0.3 gram to or from such a small amount of calcium in the extracellular fluid would obviously cause serious hyper- or hypocalcemia. However, there is a first line of defense to prevent this from occurring even before the parathyroid and calcitonin hormone feedback systems have a chance to act. This is the following mechanism:

Buffer Function of the Exchangeable Calcium in the Bones. The exchangeable calcium salts in the bones, which were discussed earlier in this chapter, are amorphous calcium phosphate compounds, probably mainly $CaHPO_4$ or some similar compound loosely bound in the bone and in reversible equilibrium with the calcium and phosphate ions in the extracellular fluid. The quantity of these salts that is available for exchange is about 0.5 to 1 per cent of the total calcium salts of the bone, a total of 5 to 10 grams of calcium. Because of the ease of deposition of these exchangeable salts and their ease of resolubility, an increase in the concentrations of extracellular fluid calcium and phosphate ions above normal causes immediate deposition of exchangeable salt. Conversely, a decrease in these concentrations causes immediate absorption of exchangeable salt. This reaction is so rapid that a single passage through a bone of blood containing a high concentration of calcium will remove almost all the excess calcium. This rapid effect results from the fact that the amorphous bone crystals are extremely small, and their total surface area exposed to the fluids of the bone is perhaps an acre or more. Also, about 5 per cent of all the blood flows through the bones each minute—that is, about 1 per cent of all the extracellular fluid each minute. Therefore, about half of any excess calcium that appears in the extracellular fluid is removed by this buffer function of the bones in approximately 70 minutes.

In addition to the buffer function of the bones, mitochondria of many of the tissues of the body, especially of the liver and intestine, also contain a reasonable amount of exchangeable calcium that provides an additional buffer system for helping maintain constancy of the extracellular fluid calcium ion concentration.

Hormonal Control of Calcium Ion Concentration, the Second Line of Defense. At the same time that the exchangeable calcium mechanism in the bones is "buffering" the calcium in the extracellular fluid, both the parathyroid and calcitonin hormonal systems are also beginning to act. Within 3 to 5 minutes after an acute increase in the calcium ion concentration, the rate of parathyroid hormone secretion has already decreased. As already explained, this sets into play multiple mechanisms for reducing the calcium ion concentration back toward normal. However, this is a slow mechanism. On the other hand, at the same time that parathyroid hormone decreases, calcitonin increases. In young animals, and presumably in young children as well (but probably very little in adults), the calcitonin causes rapid deposition of calcium in the bones, and perhaps also in some cells of other tissues. Therefore, in very young animals, excess calcitonin can cause a high calcium ion concentration to return to normal probably two to three times as rapidly as can be achieved by the exchangeable calcium-buffering mechanism alone.

In prolonged calcium excess or prolonged calcium deficiency, only the parathyroid hormonal mechanism seems to be really important in maintaining a normal plasma calcium ion concentration. When a person has a continuing deficiency of calcium in the diet, parathyroid hormone often can stimulate enough calcium absorption from the bones to maintain a normal plasma calcium ion concentration for a year or more; but eventually even the bones will run out of calcium. Thus, in effect, the bones are a large buffer-reservoir of calcium that can be manipulated by parathyroid hormone, but the ultimate very long-term control of calcium ion concentration resides in the roles of parathyroid hormone and vitamin D in controlling calcium absorption from the gut and calcium excretion in the urine.

PHYSIOLOGY OF PARATHYROID AND BONE DISEASES

HYPOPARATHYROIDISM

When the parathyroid glands do not secrete sufficient parathyroid hormone, the osteoclasts of the bone become almost totally inactive. As a result, bone reabsorption is so depressed that the level of calcium in the body fluids decreases. Osteoblastic activity is concomitantly decreased, but because calcium and phosphates are not being absorbed from the bone, the bone usually remains strong.

When the parathyroid glands are suddenly removed, the calcium level in the blood falls from the normal of

9.4 to 6 to 7 mg/dl within two to three days When this level is reached, the usual signs of tetany develop. Among the muscles of the body especially sensitive to tetanic spasm are the laryngeal muscles. Spasm of these obstructs respiration, which is the usual cause of death in tetany unless appropriate treatment is applied.

If all four parathyroid glands are removed from an animal and the animal is prevented from dying of respiratory spasm by appropriate supportive measures, the total lack of parathyroid hormone secretion normally causes the blood level of calcium to fall to as low as approximately 4 to 5 mg dl while at the same time the level of phosphates increases from the normal of about 4 mg.dl to approximately 12 mg dl because of decreased renal phosphate excretion.

Treatment of Hypoparathyroidism. *Parathyroid Hormone (Parathormone).* Parathyroid hormone is occasionally used for treating hypoparathyroidism However, because of the expense of this hormone, because its effect lasts a few hours at most, and because the tendency of the body to develop immune bodies against it makes it progressively less and less active in the body, treatment of hypoparathyroidism with parathyroid hormone is rare in present-day therapy

Vitamin D and Calcium Therapy. In most patients, administration of extremely large quantities of vitamin D, to as high as 100,000 units per day, along with the administration of 1 to 2 grams of calcium, will suffice to keep the calcium ion concentration in a normal range At times it might be necessary to administer 1,25-dihydroxycholecalciferol instead of the nonactivated form of vitamin D because of its much more potent and much more rapid action, but this can also cause unwanted effects, because it is sometimes difficult to prevent overactivity by this activated form of vitamin D

HYPERPARATHYROIDISM

The cause of hyperparathyroidism ordinarily is a tumor of one of the parathyroid glands; such tumors occur much more frequently in women than in men or children, probably because pregnancy, lactation, and perhaps other causes of prolonged low calcium levels, all of which stimulate the parathyroid gland, may predispose to the development of such a tumor.

In hyperparathyroidism extreme osteoclastic activity occurs in the bones, and this elevates the calcium ion concentration in the extracellular fluid while usually (but not always) depressing the concentration of phosphate ions because of increased renal excretion of phosphate.

Bone Disease in Hyperparathyroidism. Though in mild hyperparathyroidism new bone can be deposited rapidly enough to compensate for the increased osteoclastic reabsorption of bone, in severe hyperparathyroidism the osteoclastic absorption soon far outstrips osteoblastic deposition, and the bone may be eaten away almost entirely. Indeed, the reason a hyperparathyroid person comes to the doctor is often a broken bone. Radiographs of the bone show extensive decalcification and occasionally large punched-out cystic areas of the bone that are filled with osteoclasts in the form of so-called giant cell "tumors." Obviously, multiple fractures of the weakened bones can result from only slight trauma, especially where cysts develop. The cystic bone disease of hyperparathyroidism is called *osteitis fibrosa cystica.*

Osteoblastic activity in the bones also increases greatly in an attempt to form new bone as the old bone is absorbed by the osteoclastic activity As explained earlier in the chapter, when the osteoblasts become active they secrete large quantities of alkaline phosphatase Therefore, one of the important diagnostic findings in hyperparathyroidism is a very high level of plasma alkaline phosphatase.

Effects of Hypercalcemia in Hyperparathyroidism. Hyperparathyroidism can at times cause the plasma calcium level to rise to as high as 12 to 15 mg/dl and rarely even higher. The effects of such elevated calcium levels, as detailed earlier in the chapter, are depression of the central and peripheral nervous systems, muscular weakness, constipation, abdominal pain, peptic ulcer, lack of appetite, and depressed relaxation of the heart during diastole

Parathyroid Poisoning and Metastatic Calcification. When, on rare occasions, extreme quantities of parathyroid hormones are secreted, the level of calcium in the body fluids rises rapidly to very high values Even the extrocellular fluid phosphate concentration also often rises markedly instead of falling as is usually the case, probably because the kidneys cannot excrete rapidly enough all the phosphate being absorbed from the bone Therefore, the calcium and phosphate in the body fluids become greatly supersaturated even for the deposition of calcium phosphate ($CaHPO_4$) crystals Therefore, these crystals begin to deposit in the alveoli of the lungs, in the tubules of the kidneys, in the thyroid gland, in the acid-producing area of the stomach mucosa, and in the walls of the arteries throughout the body This extensive *metastatic* deposition of calcium phosphate can develop within a few days

Ordinarily, the level of calcium in the blood must rise above 17 mg/dl before there is danger of parathyroid poisoning, but once such elevation develops along with some concurrent elevation of phosphate, death can occur in only a few days

Formation of Kidney Stones in Hyperparathyroidism. Most patients with mild hyperparathyroidism show few signs of bone disease and few general abnormalities as a result of elevated calcium, but nevertheless do have an extreme tendency to form kidney stones The reason for this is that all the excess calcium and phosphate absorbed from the intestines or mobilized from the bones in hyperparathyroidism is excreted by the kidneys, causing proportionate increase in the concentrations of these substances in the urine. As a result, crystals of calcium phosphate tend to precipitate in the kidney, forming calcium phosphate stones Also, calcium oxalate stones develop as a result of the high level of calcium in the urine in association with normal levels of oxalate Because the solubility of most renal stones is slight in alkaline media, the tendency for formation of renal calculi is considerably greater in alkaline urine than in acid urine. For this reason, acidotic diets and acidic drugs are frequently used for treating renal calculi

Secondary Hyperparathyroidism. Because a low level of calcium ions in the body fluids directly increases the secretion of parathyroid hormone, any factor that causes a low level of calcium initiates the condition known as secondary hyperparathyroidism. This may result from low calcium diet, pregnancy, lactation, rickets, or osteomalacia. The hyperplasia of the parathyroid glands is a corrective measure for maintaining the level of calcium in the extracellular fluids at a nearly normal value.

RICKETS

Rickets occurs mainly in children as a result of calcium or phosphate deficiency in the extracellular fluid Ordinarily, rickets is due to lack of vitamin D rather than to lack of calcium or phosphate in the diet. If the child is properly exposed to sunlight, the 7-dehydrocholesterol in the skin becomes activated by the ultraviolet rays and forms vitamin D_3, which prevents rickets by promoting calcium and phosphate absorption from the intestines, as discussed earlier in the chapter.

Children who remain indoors through the winter in general do not receive adequate quantities of vitamin D without some supplementary therapy in the diet. Rickets tends to occur especially in the spring months because vitamin D formed during the preceding summer is stored in the liver and is still available for use during the early winter months. Also, calcium and phosphate absorption from the bones must take place for several months before clinical signs of rickets become apparent.

Calcium and Phosphate in the Blood of Patients with Rickets. Ordinarily, the level of calcium in the blood in rickets is only slightly depressed, but the level of phosphate is greatly depressed. This is because the parathyroid glands prevent the calcium level from falling by promoting bone absorption every time the calcium level begins to fall On the other hand, there is not a good regulatory system for preventing a falling level of phosphate, and the increased parathyroid activity actually increases the excretion of phosphates in the urine.

Effect of Rickets on the Bones. During prolonged deficiency of calcium and phosphate in the body fluids, a resulting very marked increase in parathyroid hormone secretion protects the body against hypocalcemia by causing osteoclastic absorption of the bone; this in turn causes the bone to become progressively weaker and imposes marked physical stress on the bone, resulting in rapid osteoblastic activity The osteoblasts lay down large quantities of osteoid which does not become calcified because of insufficient calcium and phosphate ions. Consequently, the newly formed, uncalcified, and very weak osteoid gradually takes the place of other bone that is being reabsorbed

Tetany in Rickets. In the early stages of rickets, tetany almost never occurs because the parathyroid glands continually stimulate osteoclastic absorption of bone and therefore maintain an almost normal level of calcium in the extracellular fluid However, when the bones become exhausted of calcium, the level of calcium may fall rapidly As the blood level of calcium falls below 7 mg/dl the usual signs of tetany develop, and the child may die of tetanic respiratory spasm unless intravenous calcium is administered, which relieves the tetany immediately

Treatment. The treatment of rickets, obviously, depends on supplying adequate calcium and phosphate in the diet and also on administering adequate amounts of vitamin D. If vitamin D is not administered along with the calcium, little calcium and phosphate are absorbed from the gut.

Osteomalacia

Osteomalacia is rickets in adults and is frequently called "adult rickets."

Normal adults rarely have serious *dietary* deficiency of vitamin D or calcium because large quantities of calcium are not needed for bone growth as in children. However, serious deficiency of vitamin D and calcium occasionally occurs as a result of steatorrhea (failure to absorb fat), for vitamin D is fat soluble, and calcium tends to form insoluble soaps with fat; consequently, in steatorrhea both vitamin D and calcium tend to pass into the feces. Under these conditions an adult occasionally has such poor calcium and phosphate absorption that adult rickets can occur, though this almost never proceeds to the stage of tetany—but very often is a cause of severe bone disability.

Osteomalacia and Rickets Caused by Renal Disease. "Renal rickets" is a type of osteomalacia resulting from prolonged kidney damage. The cause of this condition is mainly failure of the damaged kidneys to form 1,25-dihydroxycholecalciferol, the active form of vitamin D In patients whose kidneys have been completely removed or destroyed and who are being treated by hemodialysis, the problem of renal rickets is often a very severe one.

Another type of renal disease that leads to rickets and osteomalacia is *congenital hypophosphatemia* resulting from congenitally reduced reabsorption of phosphates by the renal tubules This type of rickets must be treated with phosphate compounds instead of calcium and vitamin D, and it is called *vitamin D-resistant* rickets.

OSTEOPOROSIS

Osteoporosis, the most common of all bone diseases in adults and especially in old age, is a different disease from osteomalacia and rickets, for it results from diminished organic matrix rather than abnormal bone calcification Usually, in osteoporosis the osteoblastic activity in the bone is less than normal, and consequently the rate of bone deposition is depressed. But occasionally, as in hyperparathyroidism, the cause of the diminished bone is excess osteoclastic activity

The many common causes of osteoporosis are (1) lack of physical stress on the bones because of inactivity; (2) malnutrition to the extent that sufficient protein matrix cannot be formed; (3) lack of vitamin C, which is necessary for the secretion of intercellular substances by all cells, including the osteoblasts, (4) postmenopausal lack of estrogen secretion, for estrogens have an osteoblast-stimulating activity; (5) old age, in which many of the protein anabolic functions are poor anyway so that bone matrix cannot be deposited satisfactorily; (6) Cushing's disease, because massive quantities of glucocorticoids cause decreased deposition of protein throughout the body, cause increased catabolism of protein, and also have the specific effect of depressing osteoblastic activity, and (7) acromegaly, possibly because of lack of sex hormones, excess of adrenocortical hormones, and often lack of insulin because of the diabetogenic effect of growth hormone. Thus, many different diseases of protein metabolism can cause osteoporosis.

PHYSIOLOGY OF THE TEETH

The teeth cut, grind, and mix the food eaten To perform these functions the jaws have powerful muscles capable of providing an occlusive force between the front teeth of as much as 50 to 100 pounds and as much as 150 to 200 pounds for the jaw teeth Also, the upper and lower teeth are provided with projections and facets which interdigitate so that each set of teeth fits with

the other. This fitting is called *occlusion*, and it allows even small particles of food to be caught and ground between the tooth surfaces.

FUNCTION OF THE DIFFERENT PARTS OF THE TEETH

Figure 79–11 illustrates a sagittol section of a tooth, showing its major functional parts the *enamel, dentine, cementum*, and *pulp*. The tooth can also be divided into the *crown*, which is the portion that protrudes out of the gum into the mouth, and the *root*, which is the portion that protrudes into the bony socket of the jaw. The collar between the crown and the root where the tooth is surrounded by the gum is called the *neck*

Dentine. The main body of the tooth is composed of dentine, which has a strong, bony structure Dentine is made up principally of hydroxyapatite crystals similar to those in the bone, but much more dense These are embedded in a strong meshwork of collagen fibers In other words, the principal constituents of dentine are very much the same as those of bone The major difference is its histological organization, for dentine does not contain any osteoblasts, osteocytes, osteoclasts, or spaces for blood vessels or nerves Instead, it is deposited and nourished by a layer of cells called *odontoblasts*, which line its inner surface along the wall of the pulp cavity

The calcium salts in dentine make it extremely resistant to compressional forces, while the collagen fibers make it tough and resistant to tensional forces that might result when the teeth are struck by solid objects

Enamel. The outer surface of the tooth is covered by a layer of enamel that is formed prior to eruption of the tooth by special epithelial cells called *ameloblasts* Once the tooth has erupted, no more enamel is formed Enamel is composed of very large and very dense crystals of hydroxyapatite with adsorbed carbonate, magnesium, sodium, potassium, and other ions embedded in a fine meshwork of strong and almost completely insoluble protein fibers that are similar in physical characteristics (but not chemically identical) to the keratin of hair The crystalline structure of the salts makes the enamel extremely hard, much harder than the dentine. Also, the special protein fiber meshwork, though comprising only about 1 per cent of the enamel mass, nevertheless makes enamel very resistant to acids, enzymes, and other corrosive agents because this protein is one of the most insoluble and resistant proteins known.

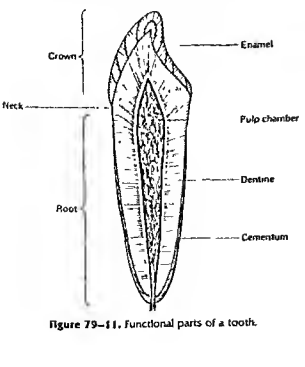


Figure 79–11. Functional parts of a tooth.

Cementum. Cementum is a bony substance secreted by cells of the *periodontal membrane*, which lines the tooth socket Many collagen fibers pass directly from the bone of the jaw, through the periodontal membrane, and then into the cementum. These collagen fibers and the cementum hold the tooth in place When the teeth are exposed to excessive strain, the layer of cementum becomes thicker and stronger. Also, it increases in thickness and strength with age, causing the teeth to become progressively more firmly seated in the jaws as one reaches adulthood and older

Pulp. The inside of each tooth is filled with pulp, which in turn is composed of connective tissue with an abundant supply of nerve fibers, blood vessels, and lymphatics The cells lining the surface of the pulp cavity are the odontoblasts, which, during the formative years of the tooth, lay down the dentine but at the same time encroach more and more on the pulp cavity, making it smaller. In later life the dentine stops growing and the pulp cavity remains essentially constant in size. However, the odontoblasts are still viable and send projections into small *dentinal tubules* that penetrate all the way through the dentine; these are of importance for providing nutrition and probably for exchange of calcium, phosphate, and other minerals.

DENTITION

Each human being and most other mammals develop two sets of teeth during a lifetime. The first teeth are called the *deciduous teeth*, or *milk teeth*, and they number 20 in the human being. These erupt between the seventh month and second year of life, and they last until the sixth to the thirteenth year. After each deciduous tooth is lost, a permanent tooth replaces it, and an additional 8 to 12 molars appear posteriorly in the jaw, making the total number of permanent teeth 28 to 32, depending on whether the four *wisdom teeth* finally appear, which does not occur in everyone.

Formation of the Teeth. Figure 79–12 illustrates the formation and eruption of teeth. Figure 79–12A shows invagination of the oral epithelium into the *dental lamina;* this is followed by the development of a tooth-producing organ. The epithelial cells above form ameloblasts, which form the enamel on the outside of the tooth. The epithelial cells below invaginate upward to form a pulp cavity and also to form the odontoblasts that secrete dentine. Thus, enamel is formed on the outside of the tooth, and dentine is formed on the inside, giving rise to an early tooth as illustrated in 79–12B.

Eruption of Teeth. During early childhood, the teeth begin to protrude upward from the jaw bone through the oral epithelium into the mouth. The cause of "eruption" is unknown, though several theories have been offered in an attempt to explain this phenomenon. The most likely theory is that growth of the tooth root as well as of the bone underneath the tooth progressively shoves the tooth forward.

Development of the Permanent Teeth. During embryonic life, a tooth-forming organ also develops in the dental lamina for each permanent tooth that will be needed after the deciduous teeth are gone. These tooth-

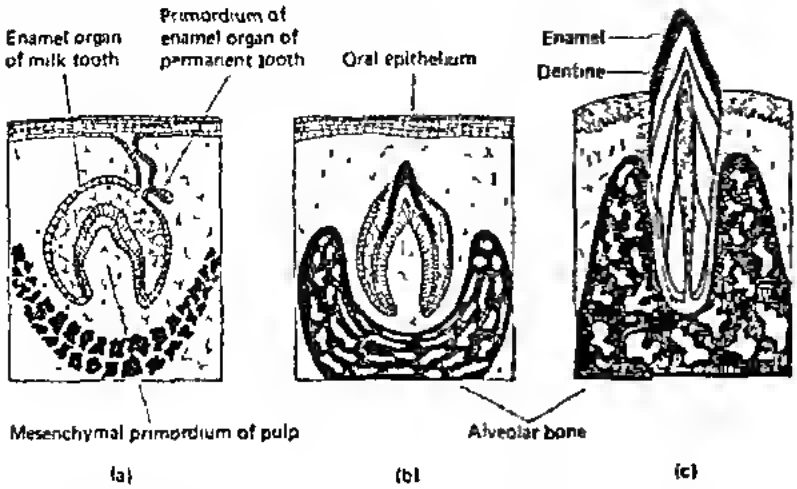


Figure 79–12. *A*, Primordial tooth organ *B*, The developing tooth *C*, The erupting tooth

producing organs slowly form the permanent teeth throughout the first 6 to 20 years of life When each permanent tooth becomes fully formed, it, like the deciduous tooth, pushes upward through the bone of the jaw In so doing it erodes the root of the deciduous tooth and eventually causes it to loosen and fall out. Soon thereafter, the permanent tooth erupts to take the place of the original one

Metabolic Factors in Development of the Teeth. The rate of development and the speed of eruption of teeth can be accelerated by both thyroid and growth hormones Also, the deposition of salts in the early forming teeth is affected considerably by various factors of metabolism, such as the availability of calcium and phosphate in the diet, the amount of vitamin D present, and the rate of parathyroid hormone secretion When all these factors are normal, the dentine and enamel will be correspondingly healthy, but, when they are deficient, the calcification of the teeth also may be defective so that the teeth will be abnormal throughout life.

MINERAL EXCHANGE IN TEETH

The salts of teeth, like those of bone, are composed basically of hydroxyapatite with adsorbed carbonates and various cations bound together in a hard crystalline substance. Also, new salts are constantly being deposited while old salts are being reabsorbed from the teeth, as also occurs in bone However, experiments indicate that deposition and reabsorption occur mainly in the dentine and cementum, while very little occurs in the enamel And most of that which does occur in the enamel occurs by exchange of minerals with the saliva instead of with the fluids of the pulp cavity. The rate of absorption and deposition of minerals in the cementum is approximately equal to that in the surrounding bone of the jaw, while the rate of deposition and absorption of minerals in the dentine is only one third that of bone. The cementum has characteristics almost identical with those of usual bone, including the presence of osteoblasts and osteoclasts, while dentine does not have these characteristics, as was explained above; this difference undoubtedly explains the different rates of mineral exchange.

The mechanism by which minerals are deposited and reabsorbed from the dentine is not clear. It is probable that the small processes of the odontoblasts that protrude into the tubules of the dentine are capable of absorbing salts and then of providing new salts to take the place of the old

In summary, rapid mineral exchange occurs in the dentine and cementum of teeth, though the mechanism of this exchange in dentine is unclear. On the other hand, enamel exhibits extremely slow mineral exchange so that it maintains most of its original mineral complement throughout life

DENTAL ABNORMALITIES

The two most common dental abnormalities are *caries* and *malocclusion* Caries means erosions of the teeth, whereas malocclusion means failure of the projections of the upper and lower teeth to interdigitate properly.

Caries, and the Role of Fluorine. It is generally agreed by all research investigators of dental caries that caries results from the action of bacteria on the teeth, the most common of which is *Streptococcus mutans* The first event in the development of caries is the deposit of *plaque*, a film of precipitated products of saliva and food, on the teeth. Large numbers of bacteria inhabit this plaque and are readily available to cause caries However, these bacteria depend to a great extent on carbohydrates for their food When carbohydrates are available, their metabolic systems are strongly activated and they also multiply. In addition, they form acids, particularly lactic acid, and proteolytic enzymes The acids are the major culprit in the causation of caries, because the calcium salts of teeth are slowly dissolved in a highly acid medium And once the salts have become absorbed, the remaining organic matrix is rapidly digested by the proteolytic enzymes.

Enamel is far more resistant to demineralization by acids than is dentine, primarily because the crystals of enamel are very dense and also are about 200 times as large in volume as the dentine crystals. Therefore, the enamel of the tooth is the primary barrier to the development of caries Once the carious process has penetrated through the enamel to the dentine, it then proceeds many times as rapidly because of the high degree of solubility of the dentine salts

Because of the dependence of the caries bacteria on carbohydrates, it has frequently been taught that eating a diet high in carbohydrate content will lead to excessive development of caries However, it is not the quantity

of carbohydrate ingested but instead the frequency with which it is eaten that is important. If eaten in many small parcels throughout the day, such as in the form of candy, the bacteria are supplied with their preferential metabolic substrate for many hours of the day, and the development of caries is extreme If eaten in large amounts only at mealtimes, the extensiveness of the caries is greatly reduced.

Some teeth are more resistant to caries than others. Studies show that teeth formed in children who drink water containing small amounts of fluorine develop enamel that is more resistant to caries than the enamel in children who drink water not containing fluorine Fluorine does not make the enamel harder than usual, but instead it displaces hydroxyl ions in the hydroxyapatite crystals which, in turn, makes the enamel several times less soluble. It is also believed that the fluorine is toxic to the bacteria as well Finally, when small pits do develop in the enamel, fluorine is believed to promote deposition of calcium phosphate to "heal" the enamel surface. Regardless of the precise means by whith fluorine protects the teeth, it is known that small amounts of fluorine deposited in enamel make teeth about three times as resistant to caries as are teeth without fluorine.

Malocclusion. Malocclusion is usually caused by a hereditary abnormality that causes the teeth of one jaw to grow to abnormal positions In malocclusion, the teeth cannot perform their normal grinding or cutting action adequately. Occasionally malocclusion also results in abnormal displacement of the lower jaw in relation to the upper jaw, causing such undesirable effects as pain in the mandibular joint or deterioration of the teeth

The orthodontist can often correct malocclusion by applying prolonged gentle pressure against the teeth with appropriate braces. The gentle pressure causes absorption of alveolar jaw bone on the compressed side of the tooth and deposition of new bone on the tensional side of the tooth. In this way the tooth gradually moves to a new position as directed by the applied pressure

REFERENCES

Al Balaghi, S, *et al* B cell differentiation factor in synovial fluid of patients with rheumatoid arthritis *Immunol Rev*, *78* 7, 1984

Albertson, D A, *et al* Subtotal parathyroidectomy versus total parathyroidectomy with autotransplantation for secondary hyperparathyroidism *Am Surg*, *51* 16, 1985

Alexander, R M, Mechanics of skeleton and tendons *In* Brooks, V B (ed) Handbook of Physiology Sec 1, Vol II Bethesda, American Physiological Society, 1981, p 17

Angus, Z S, *et al*, PTH, calcitonin, cyclic nucleotides, and the kidney *Annu Rev Physiol*, *43* 583, 1981

Avioli, L V, Calcium and osteoporosis *Annu Rev Nutr*, *4* 471, 1984

Avioli, L V, and Raisz, L G, Bone metabolism and disease *In* Bondy, P K, and Rosenberg, L E (eds) Metabolic Control and Disease, 8th Ed. Philadelphia, W B Saunders Co, 1980, p 1769

Barzel, U S (ed) Osteoporosis II. New York, Grune & Stratton, 1979

Bringhurst, F R, and Potts, J T, Jr, Calcium and phosphate distribution, turnover, and metabolic actions *In* DeGroot, L J, *et al* (eds) Endocrinology Vol 2 New York, Grune & Stratton, 1979, p 531

Copp, D H, Calcitonin Comparative endocrinology *In* DeGroot, L J, *et al* (eds) Endocrinology Vol 2 New York, Grune & Stratton, 1979, p 637

DeLuca, H F Recent advances in the metabolism of vitamin D *Annu Rev Physiol*, *43* 199, 1981

Dennis, V W, *et al* Renal handling of phosphate and calcium. *Annu Rev Physiol*, *41* 257, 1979

Fraser, D R Regulation of the metabolism of vitamin D *Physiol Rev*, *60* 551, 1980

Habener, J F Regulation of parathyroid hormone secretion and biosynthesis *Annu Rev Physiol*, *43* 211, 1981

Habener, J F, and Potts J T, Jr Diagnosis and differential diagnosis of hyperparathyroidism *In* DeGroot, L J, *et al* (eds.) Endocrinology Vol 2 New York, Grune & Stratton, 1979, p 703

Habener, J F, *et al* Parathyroid hormone Biochemical aspects of biosynthesis, secretion, action, and metabolism *Physiol Rev*, *64* 985, 1984

Ham, A W Histophysiology of Cartilage, Bone, and Joints. Philadelphia, J B Lippincott, 1979

Henry, H L, and Norman, A W Vitamin D Metabolism and biological actions *Annu Rev Nutr*, *4* 493, 1984

Jaros G G, *et al* Model of short term regulation of calcium ion concentration Simulation, *32* 193, 1979

Kelly P J Pathways of transport in bone *In* Shepherd, J T, and Abboud, F M (eds) Handbook of Physiology Sec 2, Vol III Bethesda American Physiological Society, 1983, p 371

Keutmann, H T Chemistry of parathyroid hormone *In* DeGroot, L J, *et al* (eds) Endocrinology Vol 2 New York, Grune & Stratton, 1979, p 593

Kreiger, D T (ed) Current Therapy in Endocrinology, 1983–1984 St. Louis, C V Mosby, 1983

Kumar, R Metabolism of 1,25-dihydroxyvitamin D, *Physiol Rev* *64* 478 1984

Massry, S G and Fleisch H (eds.) Renal Handling of Phosphate New York Plenum Press, 1979

Mayer G P Parathyroid hormone secretion *In* DeGroot, L J, *et al* (eds) Endocrinology Vol 2 New York, Grune & Stratton, 1979, p 607

Myers H M Fluorides and Dental Fluorosis New York, S Karger, 1978

Neilans H N, and Kimberg, D V Intestinal calcium transport Absorption secretion, and vitamin D *In* Crane, R K (ed) International Review of Physiology, Gastrointestinal Physiology III Vol 19 Baltimore, University Park Press p 227

Newman, H N Dental Plaque The Ecology of the Flora on Human Teeth Springfield, Ill, Charles C Thomas, 1980

Norman, A W Vitamin D The Calcium Homeostatic Steroid Hormone. New York, Academic Press, 1979

Parsons, J A Physiology of parathyroid hormone *In* DeGroot, L J, *et al* (eds) Endocrinology Vol 2 New York, Grune & Stratton, 1979, p 621

Pitkin R M Calcium metabolism in pregnancy and the perinatal period A review *Am J Obstet Gynecol*, *151* 99, 1985

Poole Wilson, P A, *et al* Calcium out of control. *J Mol Cell Cardiol*, *16* 175 1984

Raisz, L G and Kream, B E Hormonal control of skeletal growth *Annu Rev Physiol*, *43* 2225, 1981

Rao, G S Dietary intake and bioavailability of fluoride *Annu Rev Nutr*, *4* 115, 1984

Richmond, V L Thirty years of fluoridation A review *Am. J Clin. Nutr* *41* 129, 1985

Ryan, K J Postmenopausal estrogen use *Annu Rev Med*, *33* 171, 1982

Simmons, D J, and Kunin, A S (eds) Skeletal Research. New York, Academic Press, 1979

Smith, R. Biochemical Disorders of the Skeleton Boston, Butterworths, 1979

Stewart, A F, and Broadus, A E The regulation of renal calcium excretion An approach to hypercalciuria *Annu Rev Med*, *32* 457, 1981

Talmage, R V, and Cooper, C W; Physiology and mode of action of calcitonin *In* DeGroot, L J, *et al* (eds) Endocrinology, Vol 2. New York, Grune & Stratton, 1979, p 647

Talmage, R V, and Meyer, R A, Jr: Physiological role of parathyroid hormone *In* Greep, R O, and Astwood, G D (eds.) Handbook of Physiology Sec 7, Vol 7. Baltimore, Williams & Wilkins, 1976, p 343

Van Dop, C, and Bourne, H R. Pseudohypoparathyroidism. *Annu Rev. Med*, *34* 259, 1983.

Wasserman, R H, and Fullmer, C S: Calcium transport proteins, calcium absorption, and vitamin D *Annu Rev Physiol*, *45* 375, 1983

80

Reproductive and Hormonal Functions of the Male

The reproductive functions of the male can be divided into three major subdivisions: first, spermatogenesis, which means simply the formation of sperm, second, performance of the male sexual act; and third, regulation of male reporductive functions by the various hormones. Associated with these reproductive functions are the effects of the male sex hormones on the accessory sexual organs, on cellular metabolism, on growth, and on other functions of the body.

Physiologic Anatomy of the Male Sexual Organs. Figure 80–1 illustrates the various portions of the male reproductive system. The testis is composed of about 900 coiled *seminiferous tubules*, each about three quarters of a meter long, in which the sperm are formed. The sperm then empty into the *epididymis*, another coiled tube approximately 6 meters long. The epididymis leads into the *vas deferens*, which enlarges into the *ampulla of the vas deferens* immediately before the vas enters the body of the *prostate gland*. A *seminal vesicle*, one located on each side of the prostate, empties into the prostatic end of the ampulla, and the contents from both the ampulla and the seminal vesicle pass into an *ejaculatory duct* leading through the body of the prostate gland to empty into the *internal urethra*. *Prostatic ducts* in turn empty from the prostate gland into the ejaculatory duct. Finally, the *urethra* is the last connecting link from the testis to the exterior. The urethra is supplied with mucus derived from a large number of minute *urethral glands* located along its entire extent and even more so from bilateral *bulbourethral glands* located near the origin of the urethra.

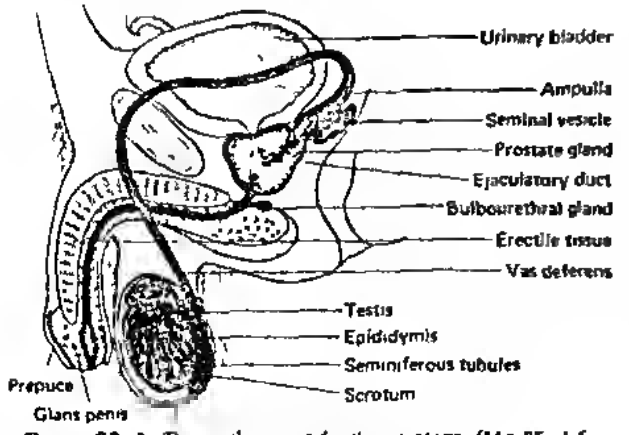


Figure 80–1. The male reproductive system. (Modified from Bloom and Fawcett. A Textbook of Histology. 10th Ed. Philadelphia, W. B. Saunders Company, 1975.)

SPERMATOGENESIS

Spermatogenesis occurs in all the seminiferous tubules during active sexual life, beginning at an average age of 13 years as the result of stimulation by anterior pituitary gonadotropic hormones and continuing throughout the remainder of life.

THE STEPS OF SPERMATOGENESIS

The seminiferous tubules, a cross-section of one of which is illustrated in Figure 80–2A, contain large numbers of germinal epithelial cells called *spermatogonia*, located in two to three layers along the outer border of the tubular epithelium. These continually proliferate to replenish themselves, and a portion of them differentiate through definite stages of development to form sperm, as shown in Figure 80–2B.

In the first stage of spermatogenesis, primitive spermatogonia located immediately adjacent to the basement membrane of the germinal epithelium, called *type A spermatogonia*, divide to form slightly more differentiated cells, the *type B spermatogonia*. Then, after several additional divisions, these cells eventually give rise to very large *primary spermatocytes*, which are the precursors of the sperm. Over a period of several weeks, each primary spermatocyte divides to form two *secondary spermatocytes*, which then divide again to form four *spermatids*. Each of these then changes slowly over the next several weeks into a spermatozoon (a sperm) by (1) losing some of its cytoplasm, (2) reorganizing the chromatin material of its nucleus to form a compact head, and (3) collecting the remaining cytoplasm and cell membrane at one end of the cell to form a tail.

All the stages of germinal cell division, as well as the final conversion of the spermatocytes into sperm, occur in contact with the *Sertoli cells*, which have specific

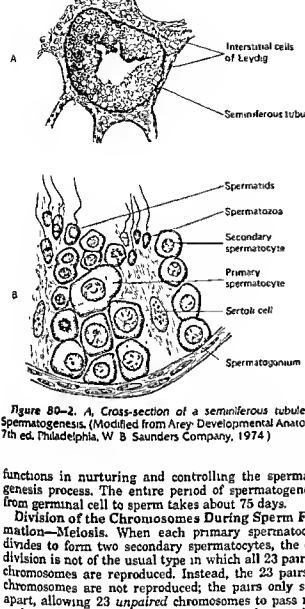


Figure 80–2. A, Cross-section of a seminiferous tubule. B, Spermatogenesis. (Modified from Arey: Developmental Anatomy 7th ed. Philadelphia, W. B. Saunders Company, 1974.)

functions in nurturing and controlling the spermatogenesis process. The entire period of spermatogenesis from germinal cell to sperm takes about 75 days.

Division of the Chromosomes During Sperm Formation—Meiosis. When each primary spermatocyte divides to form two secondary spermatocytes, the cell division is not of the usual type in which all 23 pairs of chromosomes are reproduced. Instead, the 23 pairs of chromosomes are not reproduced; the pairs only split apart, allowing 23 *unpaired* chromosomes to pass into each secondary spermatocyte. Then, as each secondary spermatocyte divides to form two spermatids, each unpaired chromosome is now reproduced, one new chromosome going to one spermatid and the other to the second spermatid. Therefore, each spermatid retains a complete set of 23 *unpaired* chromosomes, and thus each mature sperm also carries only *unpaired* chromosomes. This entire process of splitting the chromosomal pairs and forming new cells with only unpaired chromosomes is called *meiosis*.

A similar division of chromosomes occurs in the ovum. Thus, each parent makes an equal contribution to the heredity characteristics of the child.

The Sex Chromosomes. In each spermatogonium one of the 23 pairs of chromosomes carries the genetic information that determines the sex of the eventual offspring. This pair is composed of one "X" chromosome, which is called the *female chromosome*, and one "Y" chromosome, the *male chromosome*. During meiotic division the sex-determining chromosomes divide among the secondary spermatocytes so that half the sperm become *male sperm* containing the Y chromosome and the other half *female sperm* containing the X chromosome. The sex of the offspring is determined by which of these two types of sperm fertilizes the ovum. This will be discussed further in Chapter 82.

Formation of Sperm. When the spermatids are first formed, they still have the usual characteristics of epithelioid cells, but soon each spermatid begins to elongate into a spermatozoon, illustrated in Figure 80–3, composed of a *head*, and a *tail*. The head is composed of the condensed nucleus of the cell with only a very thin cytoplasmic and cell membrane layer around its surface. However, on the outside of the anterior two thirds of the head is a thick cap called the acrosome that has been formed from the Golgi apparatus. This contains a number of enzymes similar to those found in lysosomes of the typical cell, including *hyaluronidase*, which can digest proteoglycan filaments of tissues, and powerful *proteolytic enzymes*, which can digest proteins. These enzymes probably play important roles in allowing the sperm to fertilize the ovum.

The tail of the sperm has three major components: (1) a central skeleton constructed of 11 microtubules, collectively called the *axoneme*—the structure of this is similar to that of cilia described in Chapter 2; (2) a thin cell membrane covering the axoneme; and (3) a collection of mitochondria surrounding the axoneme in the proximal portion of the tail (called the *body of the tail*).

To and fro movement of the tail (flagellar movement) provides motility for the sperm. This movement is believed to result from a rhythmic longitudinal sliding motion between the anterior and posterior tubules that make up the axoneme. The energy for this process is supplied in the form of ATP synthesized by the mitochondria in the body of the tail. Normal sperm move in a straight line at a velocity of 1 to 4 mm per minute. This allows them to move through the female genital tract in quest of the ovum.

The Sertoli Cells; Their Function in Spermatogenesis

Note very carefully the very large cells, called the *Sertoli cells*, of the germinal epithelium illus-

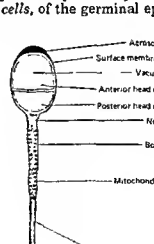


Figure 80–3. Structure of the human

trated in Figure 80–2B. These cells are massive in size and extend all the way from the basement membrane of the seminiferous epithelium to the interior of the tubule lumen, and the germinal cells and developing sperm lie in channels between the Sertoli cells. The entire process of spermatogenesis occurs in these channels in direct association with the Sertoli cells The specific functions of the Sertoli cells in spermatogenesis are:

First, the Sertoli cells provide a special environment in which the germinal cells develop They secrete a fluid that bathes the germinal cells, and they even secrete additional fluid into the lumen of the seminiferous tubule to provide appropriate nutrients for the developing and newly formed sperm

Second, the Sertoli cells play a special role in converting the spermatocytes into sperm, a process called *spermiation* During spermiation, the spermatocytes appear actually to be attached to the Sertoli cells, and it is probably digestive enzymes from the Sertoli cells that remove most of the cytoplasm from the spermatid Also, the Sertoli cells perhaps even play a physical role in shaping the head and tail of the sperm

Third, the Sertoli cells secrete several hormones that have very important functions, as follows: (1) *mullerian inhibitory factor* (MIF), secreted by the male testes during fetal development to inhibit the formation of fallopian tubes from the müllerian ducts in the male fetus, (2) *estradiol*, the principal feminizing sex hormone, which seems to be required as one of the stimulatory factors in spermatogenesis; and (3) *inhibin*, which has a feedback inhibitory effect on the anterior pituitary gland to prevent oversecretion of follicle-stimulating hormone, an important control function that we shall discuss in greater detail later

Hormonal Factors That Stimulate Spermatogenesis

We shall speak more about the role of hormones in spermatogenesis and other sexual functions later, but at this point we need to note that several hormones play absolutely essential roles in spermatogenesis. Some of these are.

(1) *Testosterone*, secreted by the Leydig cells located in the interstitium of the testis, is essential for one or more essential steps in the division of the germinal cells to form sperm, probably exciting especially the meiotic division of the primary spermatocytes to form secondary spermatocytes.

(2) *Luteinizing hormone* (LH), secreted by the anterior pituitary gland, stimulates the Leydig cells to secrete testosterone.

(3) *Follicle-stimulating hormone* (FSH), also secreted by the anterior pituitary gland, stimulates the Sertoli cells, without this stimulation, the conversion of the spermatids to sperm (the process [illegible]ation) will not occur.

(4) *Estrogens*, formed by the Sertoli cells when they are stimulated by FSH, are probably essential for the spermiation process The Sertoli cells also secrete an *androgen-binding protein* that binds both testosterone and estrogens and carries these into the fluid in the seminiferous tubule, perhaps making both these hormones available to the maturing sperm.

(5) *Growth hormone* (as well as most of the other hormones) is necessary for controlling background metabolic functions of the testes Growth hormone specifically promotes early division of the spermatogonia themselves; in its absence, as in pituitary dwarfs, spermatogenesis is severely deficient or entirely absent

Maturation of Sperm in the Epididymis

Following formation in the seminiferous tubules, the sperm require several days to pass through the 6 meter *epididymis* Sperm removed from the seminiferous tubules are completely nonmotile, and they cannot fertilize an ovum. However, after the sperm have been in the epididymis for some 18 hours to 10 days, they develop the capability of motility, even though some inhibitory factor still prevents actual motility until after ejaculation. The sperm also become capable of fertilizing the ovum, a process called *maturation* It may not be any special function of the epididymis that changes the sperm from the nonmotile state into the motile and fertile state but, instead, simply an aging process. However, the Sertoli cells and the epithelium of the epididymis secrete a special nutrient fluid containing hormones (both testosterone and estrogens), enzymes, and special nutrients that may be important or even essential for sperm maturation.

Storage of Sperm. A small quantity of sperm can be stored in the epididymis, but most sperm are stored in the vas deferens and the ampulla of the vas deferens. They can remain stored, maintaining their fertility, in the genital ducts for several months, though it is doubtful that during normal sexual activity such prolonged storage ordinarily occurs Indeed, with excessive sexual activity storage may be no longer than a few days at most.

Physiology of the Mature Sperm. The usual motile and fertile sperm are capable of flagellated movement through the fluid media at a rate of approximately 1 to 4 mm per minute. Furthermore, *normal* sperm tend to travel in a straight, rotating line rather than with a circuitous movement. The activity of sperm is greatly enhanced in neutral and slightly alkaline media as exists in the ejaculated semen, but it is greatly depressed in mildly acid media, and strong acid media can cause rapid death of sperm The activity of sperm increases greatly with increasing temperature, but so does the rate of metabolism, causing the life of

the sperm to be considerably shortened. Though sperm can live for many weeks in the genital ducts of the testes, the life of sperm in the female genital tract is only 1 to 2 days.

FUNCTION OF THE SEMINAL VESICLES

From early anatomical studies of the seminal vesicles it was erroneously believed that sperm were stored in these, whence came the name "seminal vesicles." However, these structures are only secretory glands instead of sperm storage areas

Each seminal vesicle is a tortuous, loculated tube lined with a secretory epithelium that secretes a mucoid material containing an abundance of *fructose* and other nutrient substances, as well as large quantities of *prostaglandins* and *fibrinogen*. During the process of emission each seminal vesicle empties its contents into the ejaculatory duct shortly after the vas deferens empties the sperm. This adds greatly to the bulk of the ejaculated semen, and the fructose and other substances in the seminal fluid are of considerable nutrient value for the ejaculated sperm until one of them fertilizes the ovum The prostaglandins are believed to aid fertilization in two ways (1) by reacting with the cervical mucus to make it more receptive to sperm, and (2) possibly causing reverse peristaltic contractions in the uterus and fallopian tubes to move the sperm toward the ovaries (a few sperm reach the upper end of the fallopian tubes within 5 minutes)

FUNCTION OF THE PROSTATE GLAND

The prostate gland secretes a thin, milky, alkaline fluid containing citric acid, calcium, acid phosphate, a clotting enzyme, and a profibrinolysin During emission, the capsule of the prostate gland contracts simultaneously with the contractions of the vas deferens so that the thin, milky fluid of the prostate gland adds to the bulk of the semen The alkaline characteristic of the prostatic fluid may be quite important for successful fertilization of the ovum, because the fluid of the vas deferens is relatively acidic owing to the presence of metabolic end-products of the sperm and, consequently, inhibits sperm fertility Also, the vaginal secretions of the female are acidic (pH of 3.5 to 4.0). Sperm do not become optimally motile until the pH of the surrounding fluids rises to approximately 6 to 6.5. Consequently, it is probable that prostatic fluid neutralizes the acidity of these other fluids after ejaculation and greatly enhances the motility and fertility of the sperm.

SEMEN

Semen, which is ejaculated during the male sexual act, is composed of the fluids from the vas deferens, from the seminal vesicles, from the prostate gland, and from the mucous glands, especially the bulbo-urethral glands. The major bulk of the semen is seminal vesicle fluid (about 60 per cent), which is the last to be ejaculated and serves to wash the sperm out of the ejaculatory duct and urethra. The average pH of the combined semen is approximately 7.5, the alkaline prostatic fluid having neutralized the mild acidity of the other portions of the semen. The prostatic fluid gives the semen a milky appearance, while fluid from the seminal vesicles and from the mucous glands gives the semen a mucoid consistency. Also, the clotting enzyme of the prostatic fluid causes the fibrinogen of the seminal vesicle fluid to form a weak coagulum, which then dissolutes during the next 15 to 20 minutes because of lysis by fibrinolysin formed from the prostatic profibrinolysin In the early minutes after ejaculation, the sperm remain relatively immobile, possibly because of the viscosity of the coagulum However, after the coagulum dissolutes, the sperm simultaneously become highly motile

Though sperm can live for many weeks in the male genital ducts, once they are ejaculated in the semen their maximal life span is only 24 to 48 hours at body temperature. At lowered temperatures, however, semen may be stored for several weeks; and when frozen at temperatures below −100° C, sperm have been preserved for years

Capacitation. In lower animals, sperm removed from the male genital ducts are not capable of penetrating the layer of granulosa cells that covers the ovum at the time of ovulation, but they develop this capability within a few hours after being deposited in the female genital tract This phenomenon is called *capacitation* It is believed to result from the following effects The acrosomes of the sperm contain extremely powerful hydrolytic and proteolytic enzymes that could literally destroy the male genital tract if all the sperm should release these enzymes prematurely Fortunately, the fluids in the seminiferous tubules, in the epididymis, and in the vas deferens all contain minute floating vesicles filled with large amounts of cholesterol. This cholesterol is continually donated to the cellular membrane covering the acrosome, toughening this membrane and preventing release of the enzymes. After ejaculation, the sperm swim away from the vesicles and gradually lose their excess cholesterol during the next few hours. In so doing they become "capacitated" so that the acrosomes can now release the enzymes that allow a sperm to enter the ovum.

ABNORMAL SPERMATOGENESIS AND MALE FERTILITY

The seminiferous tubular epithelium can be destroyed by a number of different diseases For instance, bilateral orchitis resulting from mumps causes sterility in a large percentage of afflicted males Also, many male infants

nerves to the sacral regions of the cord. In turn, signals from the sacral cord further excite the rhythmic contraction of the internal genital organs and also cause contraction of the ischiocavernosus and bulbocavernosus muscles that compress the bases of the penile erectile tissue. These effects together cause rhythmic, wavelike increases in pressure in the genital ducts and urethra, which "ejaculate" the semen from the urethra to the exterior. This process is called *ejaculation* At the same time, rhythmic contractions of the pelvic muscles and even of some of the muscles of the body trunk cause thrusting movements of the pelvis and penis, which also help propel the semen into the deepest recesses of the vagina and perhaps even through the cervix into the uterus.

This entire period of emission and ejaculation is called the *male orgasm.* At its termination, the male sexual excitement disappears almost entirely within one to two minutes, and erection ceases.

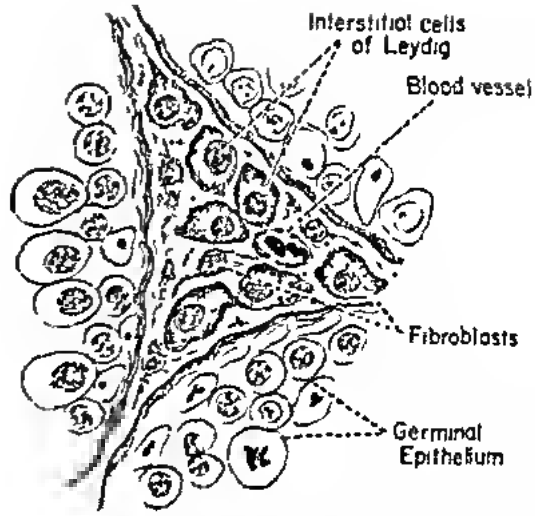


Figure 80–6. Interstitial cells of Leydig located in the interstices between the seminiferous tubules

TESTOSTERONE AND OTHER MALE SEX HORMONES

SECRETION, METABOLISM, AND CHEMISTRY OF THE MALE SEX HORMONE

Secretion of Testosterone by the Interstitial Cells of Leydig in the Testes. The testes secrete several male sex hormones, which are collectively called *androgens*, including *testosterone, dihydrotestosterone*, and *androstenedione*. However, testosterone is so much more abundant than the others that one can consider it to be the significant testicular hormone, though as we shall see, much if not most of the testosterone is converted into the more active hormone dihydrotesterone in the target tissues

Testosterone is formed by the *interstitial cells of Leydig*, which lie in the interstices between the seminiferous tubules and constitute about 20 per cent of the mass of the adult testes, as illustrated in Figure 80–6 Leydig cells are almost nonexistent in the testes during childhood, when the testes secrete almost no testosterone, but they are numerous in the newborn male infant and also in the adult male anytime after puberty; at both these times the testes secrete large quantities of testosterone. Furthermore, when tumors develop from the interstitial cells of Leydig, great quantities of testosterone are secreted Finally, when the germinal epithelium of the testes is destroyed by x-ray treatment or by excessive heat, the Leydig cells, which are less easily destroyed, continue to produce testosterone.

Secretion of Androgens Elsewhere in the Body. The term *androgen* means any steroid hormone that has masculinizing effects, including of course testosterone itself, but it also includes male sex hormones produced elsewhere in the body besides the testes For instance, the adrenal gland secretes at least five different androgens, though the total masculinizing activity of all these is normally so slight that they do not cause significant masculine characteristics even in women, except perhaps for causing growth of pubic and axillary hair. But when an adrenal tumor of the androgen-producing cells occurs, the quantity of androgenic hormones may then become great enough to cause all the usual male secondary sexual characteristics These effects were described in connection with the adrenogenital syndrome in Chapter 77

Rarely, embryonic rest cells in the ovary can develop into a tumor that produces excessive quantities of androgens in women, one such tumor is the *arrhenoblastoma* The normal ovary also produces minute quantities of androgens, but these are not significant.

Chemistry of the Androgens. All androgens are steroid compounds, as illustrated by the formulas in Figure 80–7 for *testosterone* and *dihydrotestosterone* Both in the testes and in the adrenals, the androgens can be synthesized either from cholesterol or directly from acetylcoenzyme A

Metabolism of Testosterone. After secretion by the testes, most of the testosterone becomes loosely bound with plasma albumin or a β-globulin called *gonadal steroid-binding globulin* and circulates in the blood for about 15 to 30 minutes By that time it either becomes fixed to the tissues or degraded into inactive products that are subsequently excreted

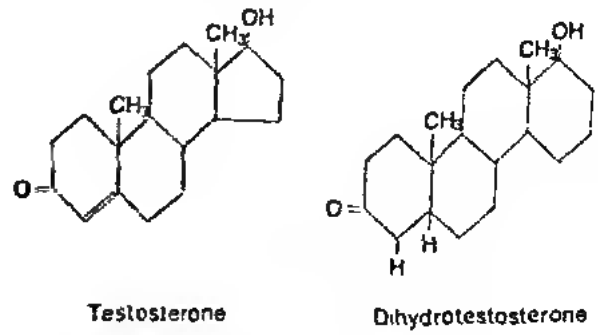


Figure 80–7. Testosterone and dihydrotesterone

Much of the testosterone that becomes fixed to the tissues is converted within the cells to *dihydrotestosterone*, and lesser quantities into 5-α-androstanediol, especially in certain target organs such as the prostate gland in the adult and in the external genitalia of the fetal male. Some actions of testosterone are dependent on this conversion, whereas other actions are not. The intracellular functions will be discussed later in the chapter.

Degradation and Excretion of Testosterone. The testosterone that does not become fixed to the tissues is rapidly converted, mainly by the liver, into *androsterone* and *dehydroepiandrosterone* and simultaneously conjugated either as glucuronides or sulfates (glucuronides, particularly). These are excreted either into the gut in the bile or into the urine.

Production of Estrogen by the Testes. In addition to testosterone, small amounts of estrogens are formed in the male (about one-fifth the amount in the nonpregnant female), and a reasonable quantity of these can be recovered from a man's urine. The functions of estrogens in the male are unknown.

The exact source of the estrogens in the male is also still doubtful, but the following are known: (a) The quantity of estrogens decreases slightly when the germinal epithelium of the seminiferous tubules is destroyed. This indicates that the seminiferous tubules, probably the Sertoli cells, synthesize estrogens in men. (b) Estrogens are formed from testosterone and androstenedione in other tissues of the body, especially the liver, probably accounting for as much as 80 per cent of the total estrogen production.

FUNCTIONS OF TESTOSTERONE

In general, testosterone is responsible for the distinguishing characteristics of the masculine body. The testes are stimulated by chorionic gonadotropin from the placenta to produce moderate quantities of testosterone during fetal development and for a weeks after birth; thereafter essentially no testosterone is produced during childhood until approximately the age of 10 to 13 years. Then testosterone production increases rapidly under the stimulus of anterior pituitary gonadotropic hormones at the onset of puberty and lasts throughout most of the remainder of life as illustrated in Figure 80–8, dwindling rapidly beyond the age of 50 to become 20 to 50 per cent of the peak value by the age of 80.

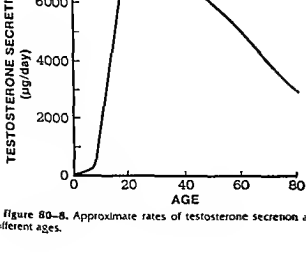


Figure 80–8. Approximate rates of testosterone secretion at different ages.

Functions of Testosterone During Fetal Development

Testosterone begins to be elaborated by the male at about the seventh week of embryonic life. Indeed, one of the major functional differences between the female and the male sex chromosome is that the male chromosome causes the newly developing genital ridge to secrete testosterone while the female chromosome causes this ridge to secrete estrogens. Injection of large quantities of male sex hormone into pregnant animals causes development of male sexual organs even though the fetus is female. Also, removal of the testes in a male fetus causes development of female sexual organs. Therefore, testosterone secreted first by the genital ridges and later by the fetal testes is responsible for the development of the male sex characteristics, including the formation of a penis and a scrotum rather than the formation of a clitoris and a vagina. Also, it causes formation of the prostate gland, the seminal vesicles, and the male genital ducts, while at the same time suppressing the formation of female genital organs.

Effect on the Descent of the Testes. The testes usually descend into the scrotum during the last two months of gestation when the testes are secreting reasonable quantities of testosterone. If a male child is born with undescended testes, administration of testosterone can still cause the testes to descend in the usual manner if the inguinal canals are large enough to allow the testes to pass. Or, administration of gonadotropic hormones, which stimulate the Leydig cells of the testes to produce testosterone, also causes the testes to descend. Thus, the stimulus for descent of the testes is testosterone, indicating again that testosterone is an important hormone for male sexual development during fetal life.

Effect of Testosterone on Development of Adult Primary and Secondary Sexual Characteristics

Testosterone secretion after puberty causes the penis, the scrotum, and the testes all to enlarge about eightfold until the age of 20 years. In addition, testosterone causes the "secondary sexual characteristics" of the male to develop at the same time, beginning at puberty and ending at maturity. These secondary sexual characteristics, in addition to the sexual organs themselves, distinguish the male from the female as follows:

Effect on the Distribution of Body Hair. Testosterone causes growth of hair (1) over the

pubis, (2) upward along the linea alba sometimes to the umbilicus and above, (3) on the face, (4) usually on the chest, and (5) less often on other regions of the body, such as the back. It also causes the hair on most other portions of the body to become more prolific.

Baldness. Testosterone decreases the growth of hair on the top of the head; a man who does not have functional testes does not become bald. However, many virile men never become bald, for baldness is a result of two factors: first, a *genetic background* for the development of baldness and, second, superimposed on this genetic background, *large quantities of androgenic hormones.* A woman who has the appropriate genetic background and who develops a long-sustained androgenic tumor becomes bald in the same manner as does a man.

Effect on the Voice. Testosterone secreted by the testes or injected into the body causes hypertrophy of the laryngeal mucosa and enlargement of the larynx. The effects cause at first a relatively discordant, "cracking" voice, but this gradually changes into the typical masculine bass voice.

Effect on the Skin, and Development of Acne. Testosterone increases the thickness of the skin over the entire body and increases the ruggedness of the subcutaneous tissues.

Testosterone increases the rate of secretion by some or perhaps all the sebaceous glands. Especially important is the excessive secretion by the sebaceous glands of the face, for oversecretion of these can result in acne. Therefore, acne is one of the most common features of adolescence when the male body is first becoming introduced to increased secretion of testosterone. After several years of testosterone secretion, the skin adapts itself to the testosterone in some way that allows it to overcome the acne.

Effect on Protein Formation and Muscular Development. One of the most important male characteristics is the development of increasing musculature following puberty, sometimes almost doubling the muscle mass. This is associated with increased protein in other parts of the body as well. Many of the changes in the skin are also due to deposition of proteins in the skin, and the changes in the voice could even result from this protein anabolic function of testosterone.

Because of the very great effect that testosterone has on the body musculature, it (or more usually a synthetic androgen instead) is widely used by athletes to improve their muscular performance. This practice is to be severely deprecated because of prolonged harmful effects of excess testosterone, as we shall discuss in Chapter 84 in relation to sports physiology. Testosterone is also frequently used in old age as a "youth hormone" to improve muscle strength and vigor.

Effect on Bone Growth and Calcium Retention. Following puberty or following prolonged injection of testosterone, the bones grow considerably in thickness and also deposit considerable additional calcium salts. Thus, testosterone increases the total quantity of bone matrix, and it also causes calcium retention. The increase in bone matrix is believed to result from the general protein anabolic function of testosterone, and the deposition of calcium salts to result from increased bone matrix.

Finally, testosterone has a specific effect on the pelvis to (a) narrow the pelvic outlet, (b) lengthen it, (c) cause a funnel-like shape instead of the broad ovoid shape of the female pelvis, and (d) greatly increase the strength of the entire pelvis for load-bearing. In the absence of testosterone, the male pelvis develops into one very similar to that of the female.

Because of the ability of testosterone to increase the size and strength of bones, testosterone is often used in old age to treat osteoporosis.

When great quantities of testosterone (or any other androgen) are secreted in the still-growing child, the rate of bone growth increases markedly, causing a spurt in total body growth as well. However, the testosterone also causes the epiphyses of the long bones to unite with the shafts of the bones at an early age in life. Therefore, despite the rapidity of growth, this early uniting of the epiphyses prevents the person from growing as tall as he would have grown had testosterone not been secreted at all. Even in normal men the final adult height is slightly less than that which would have been attained had the person been castrated prior to puberty.

Effect on Basal Metabolism. Injection of large quantities of testosterone can increase the basal metabolic rate by as much as 15 per cent, and it is believed that even the usual quantity of testosterone secreted by the testes during active sexual life increases the rate of metabolism some 5 to 10 per cent above the value that it would be were the testes not active. This increased rate of metabolism is possibly an indirect result of the effect of testosterone on protein anabolism, the increased quantity of proteins—the enzymes especially—increasing the activities of all cells.

Effect on the Red Blood Cells. When normal quantities of testosterone are injected into a castrated adult, the number of red blood cells per cubic millimeter of blood increases approximately 20 per cent. Also, the average man has about 700,000 more red blood cells per cubic millimeter than the average woman. However, this difference may be due partly to the increased metabolic rate following testosterone administration rather than to a direct effect of testosterone on red blood cell production.

Effect on Electrolyte and Water Balance. As pointed out in Chapter 77, many different steroid hormones can increase the reabsorption of sodium in the distal tubules of the kidneys. Testosterone has such an effect but only to a minor degree in comparison with the adrenal mineralocorticoids. Nevertheless, following puberty the blood and ex-

tracellular fluid volumes of the male in relation to his weight increase a slight extent.

BASIC INTRACELLULAR MECHANISM OF ACTION OF TESTOSTERONE

Probably all or almost all the effects just listed result from increased rate of protein formation in the target cells. This has been studied extensively in the prostate gland, one of the organs that is most affected by testosterone. In this gland, testosterone enters the cells within a few minutes after secretion, is there converted to *dihydrotestosterone*, and binds with a cytoplasmic "receptor protein." This combination then migrates to the nucleus where it binds with a nuclear protein and induces the DNA-RNA transcription process. Within 30 minutes RNA polymerase has become activated and the concentration of RNA begins to increase in the cells; this is followed by progressive increase in cellular protein. After several days the quantity of DNA in the gland has also increased, and there has been a simultaneous increase in the number of prostatic cells.

Therefore, it is assumed that testosterone greatly stimulates production of proteins in general, though increasing more specifically those proteins in "target" organs or tissues responsible for the development of secondary sexual characteristics.

In some tissues testosterone is converted into still another androgen, 5 α-*androstanediol*. This functions in much the same way as dihydrotestosterone, but it is considerably less potent and therefore is generally considered to be relatively unimportant.

Some important target tissues do not have appropriate enzymes in their cells to convert testosterone into dihydrotestosterone. In these tissues, testosterone functions directly, though often with less potency, to induce the formation of cellular proteins. For instance, the direct effect of testosterone is essential in the male fetus for the development of the epididymis, the vas deferens, and the seminal vesicles. The direct effect is probably also responsible for much of the effects of testosterone on the musculature, the skeleton, and the larynx and also to promote spermatogenesis.

CONTROL OF MALE SEXUAL FUNCTIONS BY THE GONADOTROPIC HORMONES—FSH AND LH

The anterior pituitary gland secretes two major gonadotropic hormones: (1) *follicle-stimulating hormone* (FSH); and (2) *luteinizing hormone* (LH). These are glycoprotein hormones that play major roles in the control of both male and female sexual function.

Regulation of Testosterone Production by LH. Testosterone is produced by the interstitial cells of Leydig only when the testes are stimulated by LH from the pituitary gland, and the quantity of testosterone secreted increases approximately in direct proportion to the amount of LH available.

Injection of purified LH into a child causes fibroblasts in the interstitial areas of the testes to develop into interstitial cells of Leydig, though mature Leydig cells are not normally found in the child's testes until after the age of approximately 10 years. Also, simultaneous administration of *prolactin* (another pituitary hormone that is closely associated with the gonadotropic hormones) greatly potentiates the effect of LH in promoting testosterone production.

Effect of Human Chorionic Gonadotropin on the Fetal Testes. During gestation the placenta secretes large quantities of *human chorionic gonodotropin*, a hormone that has almost the same properties as LH. This hormone stimulates the formation of Leydig cells in the testes of the fetus and causes testosterone secretion. As pointed out earlier in the chapter, the secretion of testosterone during fetal life is important for promoting formation of the male sexual organs.

Regulation of Pituitary Secretion of LH and FSH by the Hypothalamus

The gonadotropins, like cortocotropin and thyrotropin, are secreted by the anterior pituitary gland mainly in response to nervous activity in the hypothalamus. For instance, in the female rabbit, coitus with a male rabbit elicits nervous activity in the hypothalamus that in turn stimulates the anterior pituitary to secrete FSH and LH. These hormones then cause rapid ripening of follicles in the rabbit's ovaries, followed a few hours later by ovulation

Many other types of nervous stimuli are also known to affect gonadotropin secretion For instance, in sheep, goats, and deer, nervous stimuli in response to changes in weather and amount of light in the day increase the quantities of gonadotropins during one season of the year, the mating season, thus allowing birth of the young during an appropriate period for survival. Also, psychic stimuli can affect fertility of the male animal, as exemplified by the fact that transporting a bull under uncomfortable conditions can often cause almost complete temporary sterility. In the human being, too, it is known that various psychic stimuli feeding into the hypothalamus can cause marked excitatory or inhibitory effects on gonadotropin secretion, in this way sometimes greatly altering the degree of fertility.

Luteinizing Hormone Releasing Hormone (LHRH), the Hypothalamic Hormone That Stimulates Gonadotropin Secretion. In both the male and the female the hypothalamus controls gonadotropin secretion by way of the hypothalamic-hypophysial portal system, as was discussed in Chapter 75. Though there are two different gonadotropic hormones, luteinizing hormone and follicle-stimulating hormone, only one hypothalamic-releasing hormone has been discovered; this is *luteinizing hormone releasing hormone* (LHRH).

Though this hormone has an especially strong effect on inducing luteinizing hormone secretion by the anterior pituitary gland, it has a potent effect in causing follicle-stimulating hormone secretion as well. For this reason it is often also called *gonadotropin-releasing hormone.*

Luteinizing hormone releasing hormone plays a similar role in controlling gonadotropin secretion in the female, where the interrelationships are far more complex Therefore, its nature and its functions will be discussed in much more detail in the following chapter.

Reciprocal Inhibition of Hypothalamic-Anterior Pituitary Secretion of Gonadotropic Hormones by Testicular Hormones. *Feedback Control of Testosterone Secretion.* Injection of testosterone into either a male or a female animal strongly inhibits the secretion of luteinizing hormone but only slightly inhibits the secretion of the follicle-stimulating hormone This inhibition depends on normal function of the hypothalamus; therefore, it is quite clear that the following negative feedback control system operates continuously to control very precisely the rate of testosterone secretion

1 The hypothalamus secretes *luteinizing hormone releasing hormone*, which stimulates the anterior pituitary gland to secrete *luteinizing hormone*

2 Luteinizing hormone in turn stimulates *hyperplasia of the Leydig cells* of the testes and also stimulates production of *testosterone* by these cells

3 The testosterone in turn feeds back negatively to the hypothalamus, inhibiting production of luteinizing hormone releasing hormone This obviously limits the rate at which testosterone will be produced On the other hand, when testosterone production is too low, lack of inhibition of the hypothalamus leads to subsequent return of testosterone secretion to the normal level

Feedback Control of Spermatogenesis—Role of "Inhibin." It is known, too, that spermatogenesis by the testes inhibits the secretion of FSH Conversely, failure of spermatogenesis causes markedly increased secretion of FSH; this is especially true when the seminiferous tubules are destroyed, including destruction of the Sertoli cells in addition to the germinal cells. Therefore, it is believed that the Sertoli cells secrete a hormone that has a direct inhibitory effect mainly on the anterior pituitary gland (but perhaps slightly on the hypothalamus as well) to inhibit the secretion of FSH A glycoprotein hormone having a molecular weight between 10,000 and 30,000 and called *inhibin* has been isolated from cultured Sertoli cells and is probably responsible for most of the feedback control of FSH secretion and of spermatogenesis This feedback cycle is the following:

1 Follicle-stimulating hormone stimulates the Sertoli cells that provide nutrition for the devel- spermatozoa

2. The Sertoli cells release inhibin that in turn feeds back negatively to the anterior pituitary gland to inhibit the production of FSH. Thus, this feedback cycle maintains a constant rate of spermatogenesis, without underproduction or overproduction, that is required for male reproductive function

Puberty and Regulation of its Onset

Initiation of the onset of puberty has long been a mystery. In the earliest history of humanity, the belief was simply that the testicles "ripened" at this time. With the discovery of the gonadotropins, ripening of the anterior pituitary gland was considered responsible. Now it is known from experiments in which both testicular and pituitary tissues have been transplanted from infant animals into adult animals that both the testes and the anterior pituitary of the infant are capable of performing adult functions if appropriately stimulated Therefore, it is now certain that *during childhood the hypothalamus simply does not secrete luteinizing hormone releasing hormone (LHRH).*

For reasons not underatood, some maturation process in the brain causes the hypothalamus to begin secreting LHRH at the time of puberty. This secretion will not occur if the neuronal connections between the hypothalamua and other parts of the brain are not intact Therefore, the present belief is that the maturation process probably occurs elsewhere in the brain instead of in the hypothalamus. One suggested locus is the amygdala

The Male Adult Sexual Life and the Male Climacteric. Following puberty, gonadotropic hormones are produced by the male pituitary gland for the remainder of life, and at least some spermatogenesis usually continues until death Most men, however, begin to exhibit slowly decreasing sexual functions in their late 40s or 50s, and one study showed that the average age for terminating intersexual relations was 68, though the variation was great This decline in sexual function is related to decrease in testosterone secretion, as depicted in Figure 80–8 The decrease in male sexual function is called the *male climacteric* Occasionally, the male climacteric is associated with symptoms of hot flashes, suffocation, and psychic disorders similar to the menopausal symptoms of the female These symptoms can be abrogated by administration of testosterone, synthetic androgens, or even estrogens that are used for treatment of menopausal symptoms in the female

ABNORMALITIES OF MALE SEXUAL FUNCTION

THE PROSTATE GLAND AND ITS ABNORMALITIES

The prostate gland remains relatively small throughout childhood and begins to grow at puberty under the stimulus of testosterone This gland reaches an almost stationary size by the age of about 20 years and remains this size up to the age of approximately 50 years. At

that time in some men it begins to degenerate, along with the decreased production of testosterone by the testes. However, a benign prostatic fibroadenoma frequently develops in the prostate in older men and causes urinary obstruction This hypertrophy is not caused by testosterone.

Cancer of the prostate gland is a common cause of death, resulting in approximately 2 to 3 per cent of all male deaths.

Once cancer of the prostate gland does occur, the cancerous cells are usually stimulated to more rapid growth by testosterone and are inhibited by removal of the testes so that testosterone cannot be formed Also, prostatic cancer usually can be inhibited by administration of estrogens. Some patients who have prostatic cancer that has already metastasized to almost all the bones of the body can be successfully treated for a few months to years by removal of the testes, by estrogen therapy, or by both; following this therapy the metastases degenerate and the bones heal This treatment does not completely stop the cancer but does slow it and greatly diminishes the severe bone pain

HYPOGONADISM IN THE MALE

Hypogonadism is caused by any of several abnormalities: First, the person may be born without functional testes. Second, he may have undeveloped testes, owing to failure of the anterior pituitary to secrete gonadotropic hormones. Third, cryptorchidism (undescended testes) may occur, associated with partial or total degeneration of the testes Fourth, he may lose his testes, which is called *castration* Fifth, a rare child is born who has a genetic absence of androgen receptors in the target cells; in such a child testosterone is often formed in abundance but still has *no masculinizing effect*

When the testes are nonfunctional during fetal life or when there is genetic absence of androgen receptors in the target cells, none of the male sexual organs develop Instead, normal female organs are formed The reason for this is that the basic genetic characteristic of the fetus, whether it be male or female, is to form female sexual organs if *there are no sex hormones* But, in the presence of testosterone, formation of the female sexual organs is suppressed while instead male organs are induced.

When a boy loses his testes prior to puberty, a state of *eunuchism* ensues in which he continues to have infantile sexual characteristics throughout life The height of the adult eunuch is slightly *greater than* that of the normal man, though the bones are quite thin, the muscles are considerably weaker than those of normal man, and, obviously, the sexual organs and secondary sexual characteristics are those of a child rather than those of an adult. The voice is childlike, there is no loss of hair on the head, and the normal masculine hair distribution on the face and elsewhere *does not occur.*

When a man is castrated following puberty, some male secondary sexual characteristics revert to those of a child, and others remain of masculine character. The sexual organs regress slightly in size but not to a childlike state, and the voice regresses from the bass quality only slightly. On the other hand, there is loss of masculine hair production, loss of the thick masculine bones, and loss of the musculature of the virile male.

In the castrated adult male, sexual desires are decreased but not totally lost, provided sexual activities have been practiced previously. Erection can still occur

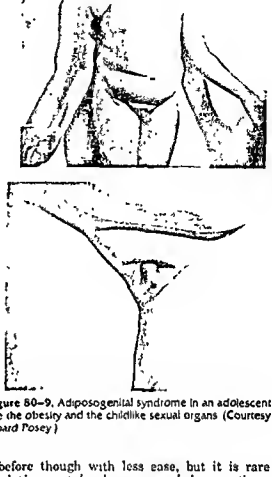


Figure 80–9. Adiposogenital syndrome in an adolescent male. Note the obesity and the childlike sexual organs (Courtesy of Dr Leonard Posey)

as before though with less ease, but it is rare that ejaculation can take place, primarily because the semen-forming organs become degenerate, and there has been a loss of the testosterone-driven psychic desire.

Some instances of hypogonadism are caused by a genetic inability of the hypothalamus to secrete normal amounts of LHRH. Often this is associated with a simultaneous abnormality of the feeding center of the hypothalamus, causing the person to greatly overeat Consequently, severe obesity occurs along with eunuchism A patient with this condition is illustrated in Figure 80–9, the condition is called *adiposogenital syndrome, Frohlich's syndrome*, or *hypothalamic eunuchism.*

TESTICULAR TUMORS AND HYPERGONADISM IN THE MALE

Interstitial Leydig cell tumors develop rarely in the testes, but when they do develop they sometimes produce as much as 100 times the normal quantities of testosterone. When such tumors develop in young children, they cause rapid growth of the musculature and bones but also early uniting of the epiphyses so that the eventual adult height actually is less than that which would have been achieved otherwise. Obviously, such interstitial cell tumors cause excessive development of the sexual organs and of the secondary sexual characteristics. In the adult male, small interstitial cell tumors are difficult to diagnose because masculine features are already present.

Much more common than the interstitial Leydig cell tumors are tumors of the germinal epithelium. Because germinal cells are capable of differentiating

any type of cell, many of these tumors contain multiple tissues, such as placental tissue, hair, teeth, bone, skin, and so forth, all found together in the same tumorous mass called a *teratoma*. Often these tumors secrete no hormones, but if a significant quantity of placental tissue develops in the tumor, it may secrete large quantities of chorionic gonadotropin that has functions very similar to those of LH Also, estrogenic hormones are frequently secreted by these tumors and cause the condition called *gynecomastia*, which means overgrowth of the breasts

THE PINEAL GLAND—ITS FUNCTION IN CONTROLLING SEASONAL FERTILITY

For as long as the pineal gland has been known to exist, to it has been ascribed a myriad of different functions, even including being the seat of the soul. However, it is known from comparative anatomy studies that the pineal gland is a vestigial remnant of what was a third eye high in the back of the head in lower animals. Many physiologists have been content with the idea that this gland is in truth merely a vestigial remnant, but others have claimed for many years that it plays important roles in the control of sexual activities and reproduction, functions that still others said were nothing more than the zealous imaginings of physiologists preoccupied with sexual delusions.

But now, after years of turmoil and dispute, it looks as though the sex advocates have at last won, and that the pineal gland does indeed play an important regulatory role in sexual and reproductive function. For, in lower animals in which the pineal gland has been removed or in which the nervous circuits to the pineal gland have been sectioned, the normal annual periods of seasonal fertility are lost. To these animals such seasonal fertility is very important because it allows a birth of the offspring in the spring and summer months when survival is most likely Unfortunately, the mechanism of this effect is still not entirely clear but it seems to be the following:

First, the pineal gland is controlled by the amount of light seen by the eyes each day For instance, in the hamster, greater than 13 hours of darkness each day activates the pineal gland, while less than that amount of darkness fails to activate it, with a very critical balance between activation and nonactivation The nervous pathway involved is this: passage of light signals from the eyes to the suprachiasmal nucleus of the hypothalamus and thence to the pineal gland activates pineal secretion

Second, the pineal gland secretes *melatonin* and several other similar substances. Either melatonin or one of the other substances then passes either by way of the blood or through the fluid of the third ventricle to the anterior pituitary gland to [illegible] gonadotropic hormone secretion.

In the presence of pineal gland secretion, the gonadotropic hormone secretion is suppressed in at least some species of animals, and the gonads become inhibited and even involuted. This is what presumably occurs during the winter months when there is more darkness. But after about four months of dysfunction, the gonadotropic hormone secretion breaks through the inhibitory effect of the pineal gland, and the gonads become functional once more, ready for a full springtime of activity.

But, does the pineal gland have a similar function for control of reproduction in man? The answer to this is still far from known. However, tumors often occur in the region of the pineal gland. Some of these are pineal tumors that secrete excessive quantities of pineal hormones, whereas others are tumors of surrounding tissue and press on the pineal gland to destroy it. Both types of tumors are often associated with serious hypo- or hypergonadal function. So, perhaps the pineal gland does play at least some role in controlling sexual drive and reproduction in man.

REFERENCES

Arimura, A Hypothalamic gonadotropin releasing hormone and reproduction *Int Rev Physiol*, *13* 1, 1977

Arnold, A P, and Gorski R A Gonadal steroid induction of structural sex differences in the central nervous system *Annu Rev Neurosci*, 7 413 1984

Bardin C W Pituitary-testicular axis *In* Yen, S S C., and Jaffe, R B (eds) Reproductive Endocrinology Philadelphia, W B Saunders Co, 1978 p 110

Bardin, C W *et al* Extracellular androgen binding proteins *Annu Rev Physiol*, *43* 189, 1981

Beckett, S D Circulation to male reproductive organs *In* Shepherd, J T and Abboud, F M (eds) Handbook of Physiology Sec 2, Vol III Rethesda, American Physiological Society, 1983, p 271

Bedford, J M Maturation, transport, and fate of spermatozoa in the epididymis *In* Greep, R O, and Astwood E B (eds.) Handbook of Physiology Sec 7, Vol 5 Baltimore, Williams & Wilkins, 1975, p 303

Binkley, S A timekeeping enzyme in the pineal gland *Sci. Am*, *240(4)* 66, 1979

Brooks, O E. Metabolic activity in the epididymis and its regulation by androgens *Physiol Rev*, 61 515, 1981

Buckner, W P, Jr Medical Readings on Human Sexuality Stanford Cal, Medical Readings Inc, 1978

Conn, P M, *et al* Gonadotropin releasing hormone Molecular and cell biology, physiology and clinical applications *Fed Proc*, *43* 2451, 1984

Davidson, J M Neurohormonal basis of male sexual behavior *Int Rev Physiol*, *13* 225, 1977

Donahoe, P K *et al* Mechanism of action of mullerian inhibiting substance *Annu Rev Physiol*, *46* 53, 1984

Eik Nes K B Biosynthesis and secretion of testicular steroids *In* Greep, R O, and Astwood, E B (eds), Handbook of Physiology Sec 7, Vol 5 Baltimore, Williams & Wilkins, 1975, p 95

Epel, D. The program of fertilization Sci Am, *237(5)* 128, 1977

Ewing L. L., *et al* Regulation of testicular function A spatial and temporal view *In* Greep, R. O (ed.) International Review of Physiology Reproductive Physiology III Vol 22 Baltimore, University Park Press, 1980, p 41

Fawcett, D W Ultrastructure and function of the Sertoli cell *In* Greep R O, and Astwood, E B (eds) Handbook of Physiology Sec 7 Vol 5 Baltimore Williams & Wilkins 1975 p 21

Fink, G. Feedback actions of target hormones on hypothalamus and pituitary, with special reference to gonadal steroids *Annu Rev Physiol*, *41* 571, 1979

Fonkalsrud, E W The Undescended Testis Chicago, Year Book Medical Publishers, 1978

Forleo R., and Pasini, W (eds) Medical Sexology Littleton, Mass PSG Publishing Co, 1979

Glover, T D Recent progress in the study of male reproductive physiology Testis stimulation, sperm formation, transport and maturation

(epididymal physiology), semen analysis, storage, and artificial insemination. *In* MTP International Review of Science Physiology Baltimore, University Park Press, 1974, p 221
Griffin, J E, and Wilson, J D The testis *In* Bondy, P K, and Rosenberg, L E (eds.) Metabolic Control and Disease, 8th Ed Philadelphia, W B, Saunders Co, 1980, p 1535
Habenicht, U F: Hormonal Regulation of Testicular Descent. New York, Springer-Verlag, 1983.
Hafez, E S E: Human Reproductive Physiology Ann Arbor, Mich, Ann Arbor Science, 1978
Hafez, E S E (ed.), Human Reproduction Conception and Contraception. Hagerstown, Md., Harper & Row, 1979
Hafez, E. S E (ed.) Descended and Cryptorchid Testis Hingham, Mass, Kluwer Boston, 1980
Hafez, E S E and Spring Mills, E (eds.) Accessory Glands of the Male Reproductive Tract. Ann Arbor, Mich, Ann Arbor Science, 1979
Hall, P F. Testicular hormones. Synthesis and control *In* DeGroot, L. J., *et al.* (eds.), Endocrinology, Vol 3 New York, Grune & Stratton, 1979, p. 1511.
Hawkins, D. F, and Elder, M G. Human Fertility Control The Theory and Practice. Boston, Butterworths, 1979
Holmes, R. L., and Fox, C A., Control of Human Reproduction New York, Academic Press, 1977.
Hull, R. W, and Pauertstein, C. J Physiology and Pathophysiology of Human Reproduction New York, John Wiley & Sons, 1978
Jaffe, R. B: Disorders of sexual development *In* Yen, S S C, and Jaffe, R B (eds.) Reproductive Endocrinology Philadelphia, W B Saunders Co., 1978, p. 271.
Janne, O. A., and Bardin, C W Androgen and antiandrogen receptor binding *Annu Rev, Physiol, 46 107, 1984*
Kelodey, R. C, *et al* Textbook of Sexual Medicine Boston, Little Brown, 1979.
Krieger, D T. (ed.), Current Therapy in Endocrinology. 1983–1984 St Louis, C. V. Mosby, 1983
Lipsett, M. B: Endocrine mechanisms of reproduction *In* Frohlich, E D (ed.) Pathophysiology, 2nd Ed Philadelphia, J B Lippincott Co 1976, p 405
Lipsett, M B, Steroid hormones. *In* Yen, S S. G, and Jaffe, R B (eds.), Reproductive Endocrinology Philadelphia, W B Saunders Co, 1978, p 50
McCann, S M., Physiology and pharmacology of *LHRH* and somatostatin. *Annu. Rev. Pharmacol Toxicol*, *22* 491, 1982.
McEwen, B S: Binding and metabolism of sex steroids by the hypothalamo-pituitary unit: Physiological implications. *Annu Rev Physiol*, *42* 97, 1980.
Meana, A R: Biochemical effects of follicle-stimulating hormone on the testis *In* Greep, R O, and Astwood, E B (eds.) Handbook of Physiology. Sec. 7, Vol 5 Baltimore, Williams & Wilkins, 1975, p 203.
Means, A. R., *et al.*, Regulation of the testis Sertoli cell by follicle stimulating hormone *Annu Rev Physiol*, *42*,59, 1980
Mess, B, Trentini, G P, and Tuna, L *Role of the Pineal Gland* in the Regulation of Ovulation New York, International Publications Service, 1978
Money, J, and Higham, E Sexual behavior and endocrinology (normal and abnormal) *In* DeGroot, L. J, *et al* (eds.) Endocrinology, Vol. 3 New York, Grune & Stratton, 1979, p 1353
Moore, P. Y.: The central nervous system and the neuroendocrine regulation of reproduction. *In* Yen, S S C, and Jaffe, R B (eds.) Reproductive Endocrinology. Philadelphia, W. B Saunders Co, 1978 p 3
Negro, V. A (ed.) Male Reproduction and Fertility New York, Raven Press, 1983
Ochsar, K (ed.). Endocrine Correlates of Reproduction. New York, Springer-Verlag, 1984
Odell, W. D. The physiology of puberty Disorders of the pubertal process *In* DeGroot, L. J, *et al* (eds.) Endocrinology Vol 3 New York, Grune & Stratton, *1979, p 1363*
Phillips, D M Spermiogenesis. New York, Academic Press, 1974.
Prasad, M R. N., and Rajalakshmi, M Recent advances in the control of male reproductive functions *Int. Rev Physiol*, *13* 153, 1977
Qian, S-Z, and Wang, Z-G Gossypol A potential antifertility agent for males *Annu Rev Pharmacol Toxicol*, *24* 329, 1984
Reiter, R J (ed) The Pineal and Reproduction New York, S Karger, 1978
Relkin, R. The Pineal Montreal, Eden Press, 1976
Ross, G F, and Lipsett, M B. (eds.), Reproductive Endocrinology Clinics in Endocrinology and Metabolism. London, W B Saunders Co, Nov, *1978*
Roy, S, and Taneja, S L Vasectomy, Vasoclusion, and Vasanastomosis A Critical Appraisal. New Delhi, National Institute of Family Planning, 1974.
Schiavi, R C Male erectile disorders *Annu Rev Med*, *32*,509, 1981.
Segal, S J *The physiology of human reproduction. Sci Am*, *231*(3),52, 1974
Setchell, B P The Mammalian Testis N Y, Cornell University Press, 1978
Smith, K D Testicular function in the aging male *In* DeGroot, L J, *et al* (eds) Endocrinology Vol 3 New York, Grune & Stratton, 1979, p 1577
Stangel J J Fertility and Conception An Essential Guide for Childless Couples New York, New American Library, 1980
Steinberger, A, and Steinberger, E (eds) Testicular Development, Structure and Function. New York, Reven Press, 1980
Steinberger F *Hormonal control of spermatogenesis. In* DeGroot, L. J, *et al* Endocrinology Vol 3 New York, Grune & Stratton, 1979, p 1535
Steinberger, E Male infertility *In* DeGroot, L. J, *et al* (eds.), Endocrinology Vol 3 New York, Grune & Stratton, 1979, p 1567
Steinberger, E Structural consideration of the male productive system. *In* DeGroot, L J, *et al* (eds) Endocrinology Vol 3 New York, Grune & Stratton, 1979, p 1501
Steinberger, E, and Steinberger, A Hormonal control of testicular function in mammals *In* Greep R O and Astwood, E B (eds) Handbook of Physiology Sec 7 Vol 4, Part 2 Baltimore, Williams & Wilkins 1974, p 325
Steinberger E, and Steinberger, A Spermatogenic function of the testis *In* Greep, R O and Astwood, E B (eds.) Handbook of Physiology Sec 7, Vol 5 Baltimore Williams & Wilkins, 1975, p 1
Styne, D M, and Grumbach, M M Puberty in the male and female its physiology and disorders *In* Yen, S S C, and Jaffe, R B (eds) Reproductive Endocrinology, Philadelphia, W B Saunders Co, 1978, p 169
Talway, G P (ed) Recent Advances in Reproduction and Regulation of Fertility New York Elsevier North-Holland, 1979
Thomas, J A, and Singahl, R L (eds) Sex Hormone Receptors in Endocrine Organs. Baltimore, Urban & Schwarzenberg, *1980*
Tollison, C D, and Adams, H E Sexual Disorders. Treatment, Theory and Research New York, Gardner Press, 1979
Tyson, J E (eds) Symposium on Neuroendocrinology of Reproduction. Clinics in Obstetrics-Gynaecology London, W B Saunders Co, Aug, 1978
vanKeep, P A, *et al* Female and Male Climacteric Current Opinion, 1978 Baltimore, University Park Press 1978
Waites, G M H, and Gladwell, R T Physiological significance of fluid secretion in the testis and blood-testis barrier *Physiol Rev* 62 624, 1982

81

Female Physiology Before Pregnancy, and the Female Hormones

Female reproductive functions can be divided into two major phases: first, preparation of the female body for conception and gestation, and second, the period of gestation itself. The present chapter is concerned with the preparation of the female body for gestation, and the following chapter presents the physiology of pregnancy.

PHYSIOLOGIC ANATOMY OF THE FEMALE SEXUAL ORGANS

Figure 81–1 illustrates the principal organs of the human female reproductive tract, the most important of which are the *ovaries*, the *fallopian tubes*, the *uterus*, and the *vagina*. Reproduction begins with the development of ova in the ovaries. A single ovum is expelled from an ovarian follicle into the abdominal cavity in the middle of each monthly sexual cycle. This ovum then passes through one of the fallopian tubes into the uterus, and, if it has been fertilized by a sperm, it implants in the uterus where it develops into a fetus, a placenta, and fetal membranes.

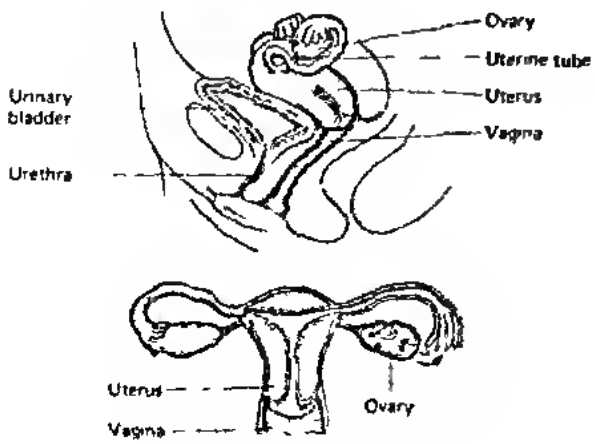


Figure 81–1. The female reproductive organs.

During fetal life the outer surface of the ovary is covered by a *germinal epithelium*, which embryologically is derived directly from the epithelium of the germinal ridges. As the fetus develops, *primordial ova* differentiate from the germinal epithelium and migrate into the substance of the ovarian cortex. Each ovum then collects around it a layer of spindle cells from the ovarian *stroma* (the supporting tissue of the ovary) and causes these to take on epithelioid characteristics; these are the *granulosa cells*. The ovum surrounded by a single layer of granulosa cells is called a *primordial follicle*.

At the 30th week of gestation the number of ova reaches about 7 million, but most of these soon degenerate so that less than 2 million are present in the two ovaries at birth, and only 300,000 to 400,000 at puberty. Then, during all the reproductive years of women, from about 13 to 50 years of age, about 450 of these follicles develop enough to expel their ova—one each month, the remainder degenerate (become *atretic*). At the end of reproductive capability, at the *menopause*, only a few primordial follicles remain in the ovaries, and even these degenerate soon thereafter.

THE FEMALE HORMONAL SYSTEM

The female hormonal system, like that of the male, consists of three different hierarchies of hormones, as follows:

1. A hypothalamic-releasing hormone, *luteinizing hormone releasing hormone* (LHRH).
2. The anterior pituitary hormones, *follicle-stimulating hormone* (FSH) and *luteinizing hormone* (LH), both of which are secreted in response to the releasing hormone from the hypothalamus.
3. The ovarian hormones, *estrogen* and *progesterone*, which are secreted by the ovaries in response to the two hormones from the anterior pituitary gland.

The various hormones are not secreted in constant amounts throughout the female monthly sexual cycle, but instead are secreted at drastically

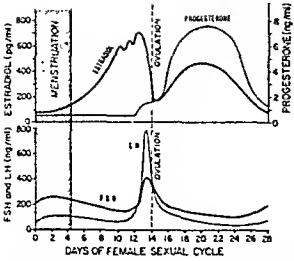


Figure 81–2. Approximate plasma concentrations of the gonadotropins and ovarian hormones during the normal female sexual cycle

differing rates during different parts of the cycle Figure 81–2 illustrates the changing concentrations of the anterior pituitary hormones, FSH and LH, and of the ovarian hormones, estradiol (estrogen) and progesterone. Although detailed measurements have not yet been made, it is possible that the hypothalamic-releasing hormone LHRH also undergoes cyclic variations.

Before we can discuss the interplay between these different hormones, it is first necessary to describe some of their specific functions and their relationships to ovarian function

THE MONTHLY OVARIAN CYCLE AND FUNCTION OF THE GONADOTROPIC HORMONES

The normal reproductive years of the female are characterized by monthly rhythmic changes in the rates of secretion of the female hormones and corresponding changes in the ovaries and sexual organs as well. This rhythmic pattern is called the *female sexual cycle* (or less accurately, the *menstrual cycle*). The duration of the cycle averages 28 days. It may be as short as 20 days or as long as 45 days even in completely normal women, though abnormal cycle length is occasionally associated with decreased fertility.

The two significant results of the female sexual cycle are: First, only a *single* mature ovum is normally released from the ovaries each month so that only a single fetus can begin to grow at a time. Second, the uterine endometrium is prepared for implantation of the fertilized ovum at the required time of the month.

The Gonadotropic Hormones

The ovarian changes during the sexual cycle depend completely on gonadotropic hormones secreted by the anterior pituitary gland. Ovaries that are not stimulated by gonadotropic hormones remain completely inactive, which is essentially the case throughout childhood when almost no gonadotropic hormones are secreted. However, at the age of about 8 years, the pituitary begins secreting progressively more gonadotropic hormones, which culminates in the initiation of monthly sexual cycles between the ages of 11 and 15 years; this culmination is called *puberty*.

(The ovaries also function during fetal life because of stimulation by a gonadotropic hormone, *chorionic gonadotropin*, secreted by the placenta. But within a few weeks after birth this stimulus is lost, and the ovaries become almost totally dormant until the prepubertal period.)

The anterior pituitary secretes two different hormones that are known to be essential for full function of the ovaries: (1) *follicle-stimulating hormone* (FSH), and (2) *luteinizing hormone* (LH). Both are small glycoproteins having molecular weights of about 30,000. The only significant effects of FSH and LH are on the testes in the male, as explained in the previous chapter, and on the ovaries in the female.

During each month of the female sexual cycle, there is a cyclic increase and decrease of FSH and LH, as illustrated in Figure 81–2. These cyclic variations in turn cause the cyclic ovarian changes, which are explained in the following sections.

Both FSH and LH stimulate their ovarian target cells by combining with highly specific FSH and LH receptors in the cell membranes. The activated receptors in turn increase both the rates of secretion of these cells and growth and proliferation of the cells Some of these stimulatory effects result from activation of the cyclic AMP second messenger system in the cell cytoplasm, which in turn excites many intracellular functions as explained in Chapter 74. However, some effects of the hormones cannot be attributed entirely to the cyclic AMP system.

Follicular Growth—Function of Follicle-Stimulating Hormone (FSH)

Figure 81–3 depicts the various stages of follicular growth in the ovaries. At birth of the female child, each of the 300,000 to 400,000 ova is surrounded by a single layer of granulosa cells, and the ovum with its granulosa cell sheath is called *a primordial follicle* as illustrated in the figure. Throughout childhood the granulosa cells are believed to provide nourishment for the ovum and also to secrete an *oocyte maturation-inhibiting factor* that keeps the ovum in its primordial state, suspended during this entire time in the first stage of meiotic division. Then, after puberty, when FSH and LH from the anterior pituitary gland begin to be secreted in large quantity, the entire ovaries

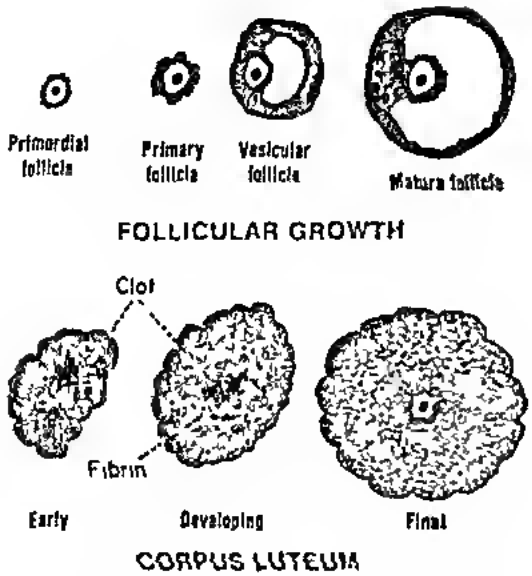


Figure 81–3. Stages of follicular growth in the ovary, showing also formation of the corpus luteum. (Modified from Arey: Developmental Anatomy. 7th ed. Philadelphia, W. B. Saunders Company, 1974.)

together with the follicles within them begin to grow.

The first stage of follicular growth is moderate enlargement of the ovum itself, which increases in diameter two to threefold. Then follows growth of additional layers of granulosa cells, and the follicle becomes known as a *primary follicle*. At least some of the development to this stage can occur even in the absence of FSH and LH, but development beyond this point is not possible without these two hormones.

The Vesicular Follicles. During the first few days after the beginning of menstruation, the concentrations of FSH and LH increase severalfold, with the increase in FSH preceding that of LH by a few days. These hormones, especially the FSH, cause accelerated growth of 6 to 12 primary follicles each month. The initial effect is rapid proliferation of the granulosa cells, but many spindle cells derived from the ovarian interstitium also collect in several layers outside the granulosa cells, giving rise to a second mass of cells called the *theca*. This is divided into two sublayers: in the *theca interna*, the cells take on epithelioid characteristics similar to those of the granulosa cells and develop the capability of secreting steroid hormones similar to the capability of the granulosa cells. The outer sublayer, the *theca externa*, is a highly vascular connective tissue capsule. This becomes the capsule of the developing follicle.

After the early proliferative phase of growth, the mass of granulosa cells secretes a *follicular fluid* that contains a high concentration of estrogen, one of the important female sex hormones that will be discussed later. The accumulation of this fluid causes an *antrum* to appear within the mass of granulosa cells, as illustrated in Figure 81–3. Once the antrum is formed, the granulosa and theca cells proliferate even more rapidly, the rate of secretion accelerates, and each of the growing follicles becomes a *vesicular follicle*.

The early growth of the primary follicle is stimulated primarily by FSH. However, the accelerated growth that leads to the vesicular follicle results from two other factors as well. First, the LH that at that time is being secreted in progressively increasing quantities by the anterior pituitary gland has a synergistic effect on the granulosa and theca cells to support the stimulatory effect of FSH, thus giving rise to far more stimulation than occurs from FSH alone. Second, the estrogen secreted into the follicle causes the granulosa cells themselves to form increasing numbers of FSH receptors, making the granulosa cells more sensitive than ever to the FSH secreted by the anterior pituitary gland. Therefore, once the vesicular follicle begins to grow, its further growth occurs very rapidly.

As the vesicular follicle enlarges, the ovum itself remains imbedded in a mass of granulosa cells located at one pole of the follicle, as illustrated in Figure 81–3. The ovum together with its surrounding granulosa cells is called the *cumulus oophorus*.

Maturation of Only One Follicle, Atresia of the Remainder. After a week or more of growth—but before ovulation occurs—one of the follicles begins to outgrow all the others; the remainder begin to involute (a process called *atresia*), and these follicles are said to become *atretic*. The cause of the atresia is unknown, but it has been postulated to be the following. The one follicle that becomes more highly developed than the others also secretes more estrogen, and this causes a positive feedback effect that (a) enhances granulosa cell proliferation, and (b) enhances the number of FSH receptors on these granulosa cells, thus increasing the stimulatory effect of FSH on this one follicle. However, the secreted estrogen acts on the hypothalamus to decrease the secretion of FSH and LH by the anterior pituitary gland, thus reducing the degree of stimulation of the less well-developed follicles. Therefore, the largest follicle continues to grow even without further increase in the gonadotropic hormones while all the other follicles stop growing and, indeed, involute.

This process of atresia obviously is important in that it allows only one of the follicles to grow large enough to ovulate. The single follicle reaches a size of 1 to 1.5 centimeters at the time of ovulation.

Ovulation

Ovulation in a woman who has a normal 28 day female sexual cycle occurs 14 days after the onset of menstruation.

Shortly before ovulation, the protruding outer wall of the follicle swells rapidly, and a small area in the center of the capsule, called the *stigma*,

protrudes like a nipple. In another half hour or so, fluid begins to ooze from the follicle through the stigma. About two minutes later, as the follicle becomes smaller because of loss of fluid, the stigma ruptures widely, and a more viscous fluid that has occupied the central portion of the follicle is evaginated outward into the abdomen. This viscous fluid carries with it the ovum surrounded by several thousand small granulosa cells called the *corona radiata*.

Need for Luteinizing Hormone (LH) in Ovulation—Ovulatory Surge of LH. Luteinizing hormone is necessary for final follicular growth and ovulation. Without this hormone, even though large quantities of FSH are available the follicle will not progress to the stage of ovulation.

Approximately two days before ovulation, for reasons that are not completely known at present but that will be discussed in more detail later in the chapter, the rate of secretion of LH by the anterior pituitary gland increases markedly, rising six- to tenfold and peaking about 18 hours before ovulation. FSH also increases about twofold at the same time, and the two of these hormones act synergistically to cause rapid swelling of the follicle during the several days before ovulation. The LH also has the specific effect on the granulosa cells of causing them at first to secrete less estrogen but progressively increasing amounts of progesterone. Therefore, the rate of secretion of estrogen begins to fall approximately one day prior to ovulation, while minute amounts of progesterone begin to be secreted.

It is in this environment of (a) very rapid growth of the follicle, (b) diminishing estrogen secretion after a prolonged phase of excessive estrogen secretion, and (c) beginning secretion of progesterone that ovulation occurs. Without the initial preovulatory surge of luteinizing hormone, ovulation will not take place.

Initiation of Ovulation. Figure 81–4 gives a schema for the initiation of ovulation. It shows the initiating cause to be the large quantity of luteinizing hormone secreted by the anterior pituitary gland. The luteinizing hormone in turn causes rapid secretion of follicular steroid hormones containing a small amount of progesterone for the first time. Within a few hours two events occur, both of which are necessary for ovulation: (1) The theca externa (the capsule of the follicle) begins to form proteolytic enzymes that cause dissolution of the capsular wall and consequent weakening of the wall, resulting in further swelling of the entire follicle and degeneration at the stigma. (2) Simultaneously, there is rapid growth of new blood vessels into the follicle wall, and at the same time prostaglandins (local hormones that cause vasodilation) are secreted in the follicular tissues. These two effects in turn cause plasma transudation into the follicle, which also contributes to follicle swelling. Finally, the combined follicle swelling and

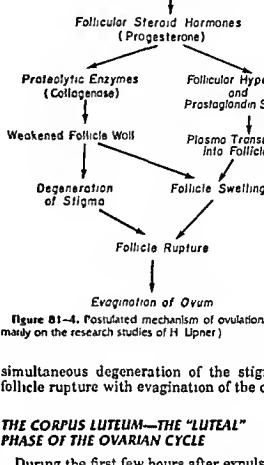


Figure 81–4. Postulated mechanism of ovulation. (Based primarily on the research studies of H. Lipner.)

simultaneous degeneration of the stigma cause follicle rupture with evagination of the ovum.

THE CORPUS LUTEUM—THE "LUTEAL" PHASE OF THE OVARIAN CYCLE

During the first few hours after expulsion of the ovum from the follicle, the remaining granulosa cells change rapidly into *lutein cells*. These grow to a diameter two or more times as large as the granulosa cells, and they become filled with lipid inclusions that give them a yellowish appearance. This process is called *luteinization*, and the total mass of cells together is called the *corpus luteum*, which is illustrated in the bottom of Figure 81–3. The lutein cells develop a very extensive smooth endoplasmic reticulum that forms large amounts of the female sex hormones progesterone and estrogen, especially progesterone. A well-developed vascular supply also grows into the corpus luteum.

Incorporated in the corpus luteum are also strands of theca cells, constituting about 20 per cent of the total mass. These cells are believed to form mainly androgens rather than female sex hormones. However, much, if not most, of the androgenic steroids are then probably converted by the granulosa cells into the female hormones.

In the normal female, the corpus luteum grows to approximately 1.5 cm, reaching this stage of development approximately seven or eight days following ovulation. After this, it begins to involute and eventually loses its secretory function, as well as its lipid characteristics, approximately 12 days following ovulation, becoming then the so-called *corpus albicans;* during the ensuing few weeks this is replaced by connective tissue.

Luteinizing Function of Luteinizing Hormone (LH). The change of granulosa cells into lutein cells is mainly dependent on the LH secreted by the anterior pituitary gland. In fact, this function gave LH its name "luteinizing." However, luteinization of the granulosa cells also depends upon extrusion of the ovum from the follicle. A yet undiscovered local hormone in the follicular fluid, called *luteinization-inhibiting factor*, seems to hold the luteinization process in check until after ovulation. Also for this reason, a corpus luteum does not develop in a follicle that does not ovulate

Secretion by the Corpus Luteum: Function of LH. The corpus luteum is a highly secretory organ, secreting large amounts of both *progesterone* and *estrogen*. Once LH (mainly that which has been secreted during the ovulatory surge) has acted on the granulosa cells to cause luteinization, the newly formed lutein cells seem to be programmed to go through a preordained sequence of (a) proliferation, (b) enlargement, and (c) secretion, then to be followed by (d) degeneration. Even in the absence of further secretion of LH by the anterior pituitary gland, this process will still occur but it will last only four to eight days. On the other hand, in the presence of LH the degree of growth of the corpus luteum is enhanced, its secretion is greater, and its life is extended. We shall see in the discussion of pregnancy in the next chapter that another hormone that has almost exactly the same properties as luteinizing hormone, *chorionic gonadotropin*, which is secreted by the placenta, can also act on the corpus luteum to prolong its life—usually maintaining it for at least the first three to four months of pregnancy.

Involution of the Corpus Luteum and Onset of the Next Ovarian Cycle. The estrogen in particular, and progesterone to a lesser extent, secreted by the corpus luteum during the luteal phase of the ovarian cycle have a strong feedback effect on the anterior pituitary gland to decrease the secretion of both FSH and LH In addition, the luteal cells also secrete moderate amounts of the hormone *inhibin*, the same as the inhibin secreted by the Sertoli cells of the male testes This hormone, too, inhibits secretion by the anterior pituitary gland, especially FSH secretion As a result, both FSH and LH in the blood fall to a very low ebb, and loss of these causes the corpus luteum to degenerate completely, a process called *involution* of the corpus luteum. This occurs at the end of almost exactly 12 days of corpus luteum life, which is approximately on the 26th day of the normal female sexual cycle, two days before menstruation begins. Now, the lack of secretion of the estrogen, progesterone, and inhibin by the corpus luteum removes the feedback inhibition of the anterior pituitary gland, allowing it again to begin secreting several times as much FSH, and a few days later increased quantities of LH as well. The FSH and LH initiate growth of new follicles to begin a ovarian cycle. At the same time, the paucity of secretion of progesterone and estrogen leads to menstruation by the uterus, as will be explained later.

SUMMARY

Approximately each 28 days, gonadotropic hormones from the anterior pituitary gland cause new follicles to begin to grow in the ovaries. One of the follicles finally ovulates at the 14th day of the cycle. During growth of the follicles, estrogen is secreted.

Following ovulation, the secretory cells of the follicle develop into a corpus luteum that secretes large quantities of the female hormones progesterone and estrogen. After another two weeks the corpus luteum degenerates, whereupon the ovarian hormones estrogen and progesterone decrease greatly and menstruation begins. A new ovarian cycle then follows.

FUNCTIONS OF THE OVARIAN HORMONES—ESTRADIOL AND PROGESTERONE

The two types of ovarian sex hormones are the *estrogens* and the *progestins*. By far the most important of the estrogens is the hormone *estradiol* and by far the most important progestin is *progesterone*. The estrogens mainly promote proliferation and growth of specific cells in the body and are responsible for development of most secondary sexual characteristics of the female. On the other hand, the progestins are concerned almost entirely with final preparation of the uterus for pregnancy and the breasts for lactation.

CHEMISTRY OF THE SEX HORMONES

The Estrogens. In the normal nonpregnant female, estrogens are secreted in major quantities only by the ovaries, though minute amounts are also secreted by the adrenal cortices In pregnancy, tremendous quantities are also secreted by the placenta, as we shall discuss in the following chapter.

Only three estrogens are present in significant quantities in the plasma of the human female: *β-estradiol*, *estrone*, and *estriol*, the formulas for which are illustrated in Figure 81–5 The principal estrogen secreted by the ovaries is β-estradiol Small amounts of estrone are also secreted, but most of this is formed in the peripheral tissues from androgens secreted by the adrenal cortices and by the ovarian thecal and stromal cells. Estriol is an oxidative product derived from both estradiol and estrone, the conversion occurring mainly in the liver.

The estrogenic potency of β-estradiol is 12 times that of estrone and 80 times that of estriol Considering these relative potencies, the total estrogenic effect of β-estradiol is usually many times that of the other two together. For this reason β-estradiol is considered to be the major estrogen, though the estrogenic effects of estrone are far from negligible.

The Progestins. By far the most important of the progestins is progesterone. However, small amounts of another progestin, 17-α-hydroxyprogesterone also are secreted along with progesterone and have essentially the same effects. Yet for practical purposes it is usually proper to consider progesterone to be the single important progestin.

In the normal nonpregnant female, progesterone is secreted in significant amounts only during the latter half of each ovarian cycle when it is secreted by the corpus luteum. Only minute amounts of progesterone appear in the plasma during the first half of the ovarian cycle, secreted approximately equally by the ovaries and the adrenal cortices. Yet, as we shall see in the following chapter, very large amounts of progesterone are also secreted by the placenta during pregnancy, especially after the fourth month of gestation.

Synthesis of the Estrogens and Progestins. Note from the chemical formulas of the estrogens and progesterone in Figure 81–5 that all these are steroids They are synthesized in the ovaries mainly from cholesterol derived from the blood, but to a slight extent also from acetyl coenzyme-A, multiple molecules of which can combine to form the appropriate steroid nucleus During synthesis, progesterone and the male sex hormone testosterone are mainly synthesized first, then, before these can leave the ovaries, almost all the testosterone and much of the progesterone are converted into estrogens Even so, about 1/15 as much testosterone is secreted into the plasma of the female by the ovaries as into the plasma of the male by the testes.

Transport of Estrogens and Progesterone in the Blood. Estrogens and progesterone are both transported in the blood bound mainly with plasma albumin, though small amounts are also bound with specific estrogen- and progesterone-binding globulins. However, the binding between these hormones and the *plasma proteins* is so loose that they are rapidly released to the tissues within 5 to 30 minutes

Fate of the Estrogens; Function of the Liver in Estrogen Degradation. The liver conjugates the estrogens to form *glucuronides and sulfates, and about one* fifth of these conjugated products are excreted in the bile while most of the remainder are excreted in the urine Also, the liver converts the potent estrogens, estradiol and estrone, into the almost totally impotent estrogen estriol Therefore, diminished liver function actually *increases* the activity of estrogens in the body, sometimes causing *hyperestrinism*

Fate of Progesterone. Within a few minutes after secretion, almost all the progesterone is degraded to other steroids that have no progesterone effect Here, as is also true with the estrogens, the liver is especially important for this metabolic degradation

The major end-product of progesterone degradation is *pregnanediol* Approximately 10 per cent of the original

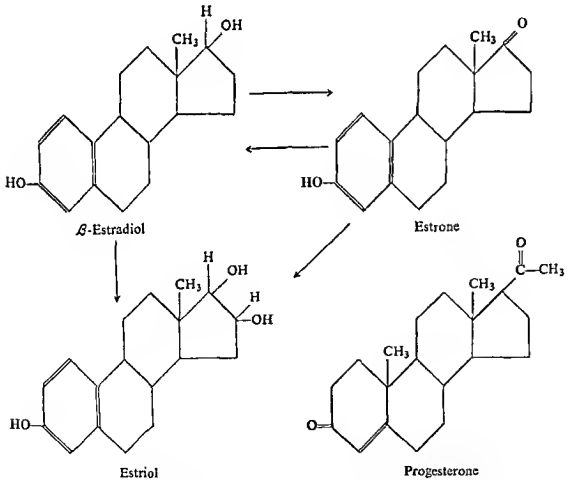


Figure 81–5. Chemical formulas of the principal female hormones.

progesterone is excreted in the urine in this form. Therefore, one can estimate the rate of progesterone formation in the body from the rate of this excretion

FUNCTIONS OF THE ESTROGENS—EFFECTS ON THE PRIMARY AND SECONDARY SEXUAL CHARACTERISTICS

The principal function of the estrogens is to cause cellular proliferation and growth of the tissues of the sexual organs and of other tissues related to reproduction.

Effect on the Uterus and External Sexual Organs. During childhood, estrogens are secreted only in small quantities, but following puberty the quantity of estrogens secreted under the influence of the pituitary gonadotropic hormones increases some 20-fold or more At this time the female sexual organs change from those of a child to those of an adult The fallopian tubes, uterus, and vagina all increase in size. Also, the external genitalia enlarge, with deposition of fat in the mons pubis and labia majora and with enlargement of the labia minora.

In addition, estrogens change the vaginal epithelium from a cuboidal into a stratified type, which is considerably more resistant to trauma and infection than is the prepubertal epithelium Infections in children, such as gonorrheal vaginitis, can actually be cured by administration of estrogens simply because of the resulting increased resistance of the vaginal epithelium.

During the few years following puberty, the size of the uterus increases two- to threefold. More important, however, than increases in size are the changes that take place in the endometrium under the influence of estrogens, for estrogens cause marked proliferation of the endometrial stroma and greatly increased development of the endometrial glands that will later be used to aid in nutrition of the implanting ovum. These effects are discussed later in the chapter in connection with the endometrial cycle.

Effect on the Fallopian Tubes. The estrogens have an effect on the mucosal lining of the fallopian tubes similar to that on the uterine endometrium They cause the glandular tissues to proliferate; and, especially important, they cause the number of ciliated epithelial cells that line the fallopian tubes to increase. Also, the activity of the cilia is considerably enhanced, these always beating toward the uterus. This helps propel the fertilized ovum toward the uterus.

Effect on the Breasts. The primordial breasts of both female and male are exactly alike, and under the influence of appropriate hormones, the masculine breast, at least during the first two decades of life, can develop sufficiently to produce milk in the same manner as the female breast.

Estrogens cause (1) development of the stromal tissues of the breasts, (2) growth of an extensive ductile system, and (3) deposition of fat in th breasts. The lobules and alveoli of the breast d velop to a slight extent, but it is progesterone ar prolactin that cause the determinative growth ar function of these structures. In summary, the e trogens initiate growth of the breasts and th breasts' milk-producing apparatus, and they a also responsible for the characteristic external a pearance of the mature female breast, but they c not complete the job of converting the breasts in milk-producing organs.

Effect on the Skeleton. Estrogens cause i creased osteoblastic activity. Therefore, at puber when the female enters her reproductive year her growth rate becomes rapid for several year However, estrogens have another potent effect c skeletal growth—that is, they cause early unitir of the epiphyses with the shafts of the long bone This effect is much stronger in the female than the similar effect of testosterone in the male. As result, growth of the female usually ceases seven years earlier than growth of the male. The fema eunuch who is completely devoid of estrogen pr duction usually grows several inches taller tha the normal mature female because her epiphyse do not unite early.

Osteoporosis of the Bones Caused by Estrogen Deficiency in Old Age. After the menopaus almost no estrogens are secreted by the ovarie This estrogen deficiency leads to (1) diminishe osteoblastic activity in the bones, (2) decrease bone matrix, and (3) decreased deposition of bon calcium and phosphate. In some women, this effe is extremely severe, and the resulting condition i *osteoporosis*, which was described in Chapter 7 Because this can greatly weaken the bones sa lead to bone fracture, especially fracture of th vertebrae, a large share of postmenopausal wome are treated continually with substitutive estro gens.

Effect on Protein Deposition. Estrogens caus a slight increase in total body protein, which i evidenced by a slight positive nitrogen balanc when estrogens are administered. This probabl results from the growth-promoting effect of estro gen on the sexual organs, the bones, and a fe other tissues of the body. The enhanced protei deposition caused by testosterone is much mor general and many times as powerful as that cause by estrogens.

Effect on Metabolism and Fat Deposition Estrogens increase the metabolic rate slightly bu only about a third as much as the male sex hor mone testosterone. They also cause deposition c increased quantities of fat in the subcutaneou tissues. As a result, the overall specific gravity c the female body, as judged by flotation in water is considerably less than that of the male body which contains more protein and less fat. In addi tion to deposition of fat in the breasts and subcu taneous tissues, estrogens cause the deposition o

fat in the buttocks and thighs that is characteristic of the feminine figure.

Effect on Hair Distribution. Estrogens do not greatly affect hair distribution. However, hair develops in the pubic region and in the axillae after puberty. Androgens formed by the adrenal glands are mainly responsible for this.

Effect on the Skin. Estrogens cause the skin to develop a texture that is soft and usually smooth but nevertheless thicker than that of the child or the female castrate. Also, estrogens cause the skin to become more vascular than normal; this effect is often associated with increased warmth of the skin, and it often results in greater bleeding of cut surfaces than is observed in men

The adrenal androgens, which are secreted in increased quantities after puberty, cause increased secretion by the axillary sweat glands and also often cause acne.

Effect on Electrolyte Balance. The chemical similarity of estrogenic hormones to adrenocortical hormones has been pointed out; and estrogens, like aldosterone and some other adrenocortical hormones, cause sodium and water retention by the kidney tubules. However, this effect of estrogens is slight and rarely of significance except in pregnancy, as discussed in the following chapter.

Intracellular Functions of Estrogens. Thus far we have discussed the gross effects of estrogens on the body. The cellular mechanism behind these effects is: Estrogens circulate in the blood for only a few minutes before they are delivered to the target cells. On entry into these cells, they combine within 10 to 15 seconds with a "receptor" protein in the cytoplasm and then, in combination with this protein, migrate to the nucleus. This in turn interacts with specific portions of the chromosomal DNA, and immediately initiates the process of transcription; therefore, RNA begins to be produced within a few minutes. In addition, over many hours new DNA also is produced, resulting eventually in division of the cell. The RNA diffuses to the cytoplasm, where it causes greatly increased protein formation and subsequently altered cellular function.

One of the principal differences between the protein anabolic effect of the estrogens and that of testosterone is that estrogen causes its effect almost exclusively in certain target organs, such as the uterus, the breasts, the skeleton, and certain fatty areas of the body, whereas testosterone has a more generalized effect throughout the body.

FUNCTIONS OF PROGESTERONE

Effect on the Uterus. By far the most important function of progesterone is *to promote secretory changes in the uterine endometrium* during the latter half of the female sexual cycle, thus preparing the uterus for implantation of the fertilized ovum. This function is discussed later in connection with the endometrial cycle of the uterus.

In addition to this effect on the endometrium, progesterone decreases the frequency of uterine contractions, thereby helping to prevent expulsion of the implanted ovum, on effect discussed in the following chapter.

Effect on the Fallopian Tubes. Progesterone also promotes secretory changes in the mucosal lining of the fallopian tubes. These secretions are necessary for nutrition of the fertilized, dividing ovum as it traverses the fallopian tube prior to implantation.

Effect on the Breasts. Progesterone promotes development of the lobules and alveoli of the breasts, causing the alveolar cells to proliferate, to enlarge, and to become secretory in nature. However, progesterone does not cause the alveoli actually to secrete milk, for, as discussed in the following chapter, milk is secreted only after the prepared breast is further stimulated by prolactin from the anterior pituitary.

Progesterone also causes the breasts to swell. Part of this swelling is due to the secretory development in the lobules and alveoli, but part also seems to result from increased fluid in the subcutaneous tissue itself.

Effect on Electrolyte Balance. Progesterone in very large quantity, like estrogens, testosterone, and adrenocortical hormones, can enhance sodium, chloride, and water reabsorption from the distal tubules of the kidney. Yet, strangely enough, progesterone more often causes increased sodium and water excretion. The cause of this is competition between progesterone and aldosterone, which probably occurs as follows: These two substances combine with the same receptor proteins in the epithelial cells of the tubules. When progesterone combines with these, aldosterone cannot combine. Yet progesterone exerts many hundred times less sodium transport effect than does aldosterone. Therefore, despite the fact that under appropriate conditions progesterone can weakly promote sodium and water retention by the renal tubules, it blocks the far more potent effect of aldosterone, thus resulting in net loss of sodium and water from the body.

Protein Catabolic Effect. Progesterone exerts a mild catabolic effect on the body's protein, similar to that of the glucocorticoids. Though in the normal sexual cycle this effect is probably not significant, it possibly is significant during pregnancy, when proteins must be mobilized for use by the fetus.

THE ENDOMETRIAL CYCLE AND MENSTRUATION

Associated with the cyclic production of estrogens and progesterone by the ovaries is an endometrial cycle operating through the following stages: first, proliferation of the uterine endometrium; second, secretory changes in the endometrium; and third, desquamation of the endome-

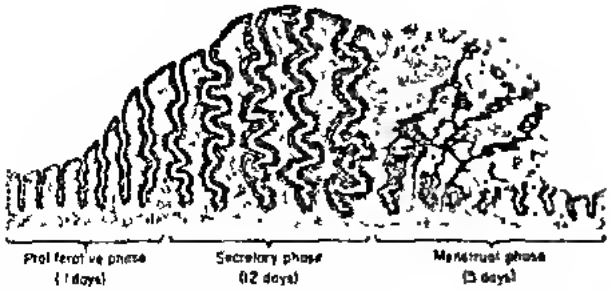


Figure 81–6. Phases of endometrial growth and menstruation during each monthly female sexual cycle

trium, which is known as *menstruation*. The various phases of the endometrial cycle are illustrated in Figure 81–6.

Proliferative Phase (Estrogen Phase) of the Endometrial Cycle. At the beginning of each monthly sexual cycle, most of the endometrium is desquamated by the process of menstruation After menstruation, only a thin layer of endometrial stroma remains at the base of the original endometrium, and the only epithelial cells that are left are those located in the remaining deep portions of the glands and crypts of the endometrium. *Under the influence of estrogens*, secreted in increasing quantities by the ovary during the first part of the ovarian cycle, the stromal cells and the epithelial cells proliferate rapidly. The endometrial surface is re-epithelialized within three to seven days after the beginning of menstruation For the first two weeks of the sexual cycle—that is, until ovulation—the endometrium increases greatly in thickness, owing to increasing numbers of stromal cells and to progressive growth of the endometrial glands and blood vessels into the endometrium. At the time of ovulation the endometrium is approximately 3 to 4 mm thick. At this time, the endometrial glands, especially those of the cervical region, secrete a thin, stringy mucus The mucous strings align themselves along the length of the cervical canal, forming channels that help guide the sperm in the proper direction into the uterus

Secretory Phase (Progestational Phase) of the Endometrial Cycle. During the latter half of the monthly cycle, progesterone as well as estrogen is secreted in large quantity by the corpus luteum. The estrogens cause slight additional cellular proliferation in the endometrium during this phase of the endometrial cycle, whereas progesterone causes marked swelling and secretory development of the endometrium. The glands increase in tortuosity; secretory substances accumulate in the glandular epithelial cells Also, the cytoplasm of the stromal cells increases, lipid and glycogen deposits increase greatly in the stromal cells, and the blood supply to the endometrium further increases in proportion to the developing secretory activity, the blood vessels becoming highly tortuous. At the end of the secretory phase the endometrium has a thickness of 5 to 6 mm.

The whole purpose of all these endometrial changes is to produce a highly secretory endometrium containing large amounts of stored nutrients that can provide appropriate conditions for implantation of a fertilized ovum during the latter half of the monthly cycle. From the time fertilization first takes place until the ovum implants, the fallopian and uterine secretions, called "uterine milk," provide nutrition for the early dividing ovum. Then, once the ovum implants in the endometrium, the trophoblastic cells on the surface of the blastocyst begin to digest the endometrium and to absorb the digested substances, thus making still far greater quantities of nutrients available to the early embryo.

Menstruation. Approximately two days before the end of the monthly cycle, the ovarian hormones, estrogens and progesterone, decrease sharply to low levels of secretion, as was illustrated in Figure 81–2, and menstruation follows

Menstruation is caused by this sudden reduction of the estrogens and progesterone at the end of the monthly ovarian cycle. The first effect is decreased stimulation of the endometrial cells by these two hormones, followed rapidly by involution of the endometrium itself to about 65 per cent of its previous thickness. During the 24 hours preceding the onset of menstruation, the tortuous blood vessels leading to the mucosal layers of the endometrium become vasospastic, presumably because of some effect of the involution, such as release of a vasoconstrictor material—possibly one of the vasoconstrictor types of prostaglandins. The vasospasm and loss of hormonal stimulation cause beginning necrosis in the endometrium, especially of blood vessels in the stratum vasculare. As a result, blood seeps into the vascular layer of the endometrium, and the hemorrhagic areas grow over a period of 24 to 36 hours. Gradually, the necrotic outer layers of the endometrium separate from the uterus at the site of the hemorrhages, until, at approximately 48 hours following the onset of menstruation, all the superficial layers of the endometrium have desquamated. The desquamated tissue and blood in the uterine cavity initiate uterine contractions that expel the uterine contents.

During normal menstruation, approximately 35 ml of blood and an additional 35 ml of serous fluid are lost. This menstrual fluid is normally nonclotting, because a *fibrinolysin* is released along with the necrotic endometrial material. However, if excessive bleeding occurs from the uterine surface, the quantity of fibrinolysin may not be sufficient to prevent clotting The presence of clots during menstruation ordinarily is clinical evidence of uterine pathology.

Within three to seven days after menstruation starts, the loss of blood ceases, for by this time the endometrium has become completely re-epithelialized.

Leukorrhea During Menstruation. During menstruation tremendous numbers of leukocytes are released along with the necrotic material and blood. It is probable that some substance liberated by the endometrial necrosis causes this outflow of leukocytes. As a result of these many leukocytes and possible other factors, the uterus is highly resistant to infection during menstruation even though the endometrial surfaces are denuded. Obviously, this is of extreme protective value.

REGULATION OF THE FEMALE MONTHLY RHYTHM—INTERPLAY BETWEEN THE OVARIAN AND HYPOTHALAMIC-PITUITARY HORMONES

Now that we have presented the major cyclic changes that occur during the female sexual cycle, we can attempt to explain the basic rhythmic mechanism that causes these cyclic variations.

Function of the Hypothalamus in the Regulation of Gonadotropin Secretion–Luteinizing Hormone Releasing Hormone (LHRH). As was pointed out in Chapter 74, secretion of most of the anterior pituitary hormones is controlled by releasing hormones formed in the hypothalamus and transported to the anterior pituitary gland by way of the hypothalamic-hypophysial portal system. In the case of the gonadotropins, at least one releasing factor, *luteinizing hormone releasing hormone (LHRH)*, is important. This has been purified and has been found to be a decapeptide having the following formula:

GLU-HIS-TRP-SER-TYR-GLY-LEU-ARG-PRO-GLY-NH_2

Though some research workers believe that another substance similar to this, called follicle-stimulating hormone releasing hormone (FSHRH), is also formed in the hypothalamus, the purified LHRH has been found to cause release not only of luteinizing hormone but also of follicle-stimulating hormone. Therefore, since there is reason to believe that this decapeptide is in reality both LHRH and FSHRH combined in the same molecule, it is often called simply *gonadotropin-releasing hormone* (GnRH).

Hypothalamic Centers for Stimulating Release of LHRH. Several different areas that profoundly influence the rate of secretion of the hypothalamic gonadotropin-releasing hormone LHRH have been found in the hypothalamus. In monkeys—and presumably in people—the mid-basal region of the hypothalamus is the area most importantly involved. In lower animals, the area around the infundibulum causes a continuous tonic secretion of gonadotropin-releasing hormone, while two other areas modulate the rate of release. These areas are (1) a center in the preoptic area that causes cyclic variation in the secretory rate and (2) a center in the posterior hypothalamus that allows the psychic attitude of the animal to enhance or decrease the secretion of gonadotropin-releasing hormone.

Effect of Psychic Factors on the Female Sexual Cycle. It is well known that the young woman on first leaving home to go to college almost as often as not experiences disruption or irregularity of the female sexual cycle. Likewise, serious stresses of almost any type can interfere with the cycle. Finally, in many lower animals no ovulation occurs at all until after copulation; the sexual excitation attendant to the sexual act initiates a sequence of events that leads first to secretion in the hypothalamus of LHRH, then to secretion of the anterior pituitary gonadotropins, and finally to ovarian secretion of the female sex hormones and to ovulation. It is these effects that are believed to be mediated through the posterior hypothalamic center for modulating the output of LHRH.

Negative Feedback Effect of Estrogen, Progesterone, and Inhibin on Secretion of Follicle-Stimulating Hormone and Luteinizing Hormone. Estrogen in small amounts and progesterone in large amounts inhibit the production of FSH and LH. Both these feedback effects seem to operate on the hypothalamus and the anterior pituitary gland as well.

In addition to the feedback effects of estrogen and progesterone, still another hormone also seems to be involved. This is *inhibin*, which is secreted along with the steroid sex hormones by the corpus luteum cells in the same way that the Sertoli cells secrete this same hormone in the male testes. This inhibin has the same effect in the female as in the male of inhibiting the secretion of FSH by the anterior pituitary gland, and LH also to a lesser extent. Therefore, it is believed that inhibin might be especially important in causing the decrease in secretion of FSH and LH toward the end of the female sexual month.

Positive Feedback Effect of Estrogen Before Ovulation—The Preovulatory Luteinizing Hormone Surge. For reasons not completely understood, the anterior pituitary gland secretes greatly increased amounts of LH for a period of about one day beginning 24 to 48 hours before ovulation. This effect is illustrated in Figure 81-2, and the figure shows a much smaller preovulatory surge of FSH as well.

Experiments have shown that infusion of estrogen into a female for a period of two to three days during the first half of the ovarian cycle will cause rapidly accelerating growth of the follicles and also rapidly accelerating secretion of ovarian estrogens. During this period the secretion of both the follicle-stimulating hormone and luteinizing hormone by the anterior pituitary gland is at first suppressed slightly. Then abruptly the

of luteinizing hormone increases about eightfold, and the secretion of follicle-stimulating hormone increases about twofold. The cause of this abrupt increase in secretion of the gonadotropins is not known. However, several possible causes are: (1) It has been suggested that estrogen at this point in the cycle has a peculiar *positive feedback effect* to stimulate pituitary secretion of the gonadotropins; this is in sharp contrast to its normal negative feedback effect that occurs during the remainder of the female monthly cycle. (2) The granulosa cells of the follicles begin to secrete increasing quantities of progesterone a day or so prior to the preovulatory LH surge, and it has been suggested that this might be the factor that stimulates the excess LH secretion.

Without the normal preovulatory surge of luteinizing hormone, ovulation will not occur.

FEEDBACK OSCILLATION OF THE HYPOTHALAMIC-PITUITARY-OVARIAN SYSTEM

Now, after discussing much of the known information about the interrelationships of the different components of the female hormonal system, we can digress from the area of proven fact into the realm of speculation and attempt to explain the feedback oscillation that controls the rhythm of the female sexual cycle. It seems to operate in approximately the following sequence of three successive events:

1. The Postovulatory Secretion of the Ovarian Hormones and Depression of Gonadotropins. The easiest part of the cycle to explain is the events that occur during the postovulatory phase—between ovulation and the beginning of menstruation. During this time the corpus luteum secretes large quantities of both progesterone and estrogen and probably the hormone inhibin as well. All these hormones together have a combined negative feedback effect on the anterior pituitary gland and the hypothalamus to cause strong suppression of both FSH and LH, decreasing these to their lowest levels at about three to four days before the onset of menstruation. These effects are illustrated in Figure 81–2.

2. The Follicular Growth Phase. Two to three days before menstruation the corpus luteum involutes, and the secretion of estrogen, progesterone, and inhibin decreases to a low ebb. This releases the hypothalamus and anterior pituitary from the feedback effect of these hormones, and a day or so later, at about the time that menstruation begins, FSH increases two- to threefold; then several days after menstruation begins, LH secretion increases about twofold as well. These hormones initiate new follicular growth and progressive increase in the secretion of estrogen, reaching a peak estrogen secretion at about 12.5 to 13 days after the onset of menstruation. During the first 11 to 12 days of this follicular growth the rates of secretion of the gonadotropins FSH and LH decrease; then comes a sudden increase in secretion of both these hormones, leading to the preovulatory surge of LH followed by ovulation.

3. Preovulatory Surge of LH and FSH; Ovulation. At approximately 11.5 to 12 days after the onset of menstruation, the decline in secretion of FSH and LH comes to an abrupt halt. It is believed that the high level of estrogens at this time (or the beginning secretion of progesterone by the follicles) causes a positive feedback effect on the hypothalamus and anterior pituitary, as explained earlier, which leads to a terrific surge of secretion of LH and to a lesser extent of FSH. Whatever the cause of this preovulatory LH and FSH surge, the LH leads to both ovulation and subsequent secretion by the corpus luteum. Thus, the hormonal system begins a new round of the cycle until the next ovulation.

Anovulatory Cycles—The Sexual Cycles at Puberty

If the preovulatory surge of luteinizing hormone is not of sufficient magnitude, ovulation will not occur, and the cycle is then said to be "anovulatory." Most of the cyclic variations of the sexual cycle continue, but they are altered in the following ways: First, lack of ovulation causes failure of development of the corpus luteum and consequently there is almost no secretion of progesterone during the latter portion of the cycle. Second, the cycle is shortened by several days, but the rhythm continues. Therefore, it is likely that progesterone is not required for maintenance of the cycle itself though it can alter its rhythm.

Anovulatory cycles are usual during the first few cycles following puberty and for several years prior to menopause, presumably because the LH surge is not potent enough at these times to cause ovulation.

PUBERTY AND MENARCHE

Puberty means the onset of adult sexual life, and menarche means the onset of menstruation. As pointed out earlier in the chapter, the period of puberty is caused by a gradual increase in gonadotropic hormone secretion by the pituitary, beginning approximately in the eighth year of life, as illustrated in Figure 81–7, and usually culminating in the onset of menstruation between the ages of 11 and 16 years.

In the female, as in the male, the infantile pituitary gland and ovaries are capable of full function if appropriately stimulated. However, as is also true in the male and for reasons not yet understood, the hypothalamus does not secrete luteinizing hormone releasing hormone during childhood. Experiments have shown that the hy-

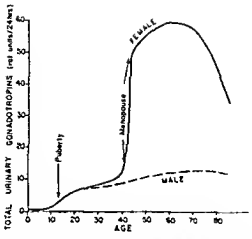


Figure 81–7. Total rates of secretion of gonadotropic hormones throughout the sexual lives of females and males showing an especially abrupt increase in gonadotropic hormones at the menopause in the female

pothalamus itself is perfectly capable of secreting this hormone, but there is lack of the appropriate signal from some other brain area to cause the secretion. Therefore, it is now believed that the onset of puberty is initiated by some maturation process occurring elsewhere in the brain besides the hypothalamus, perhaps somewhere in the limbic system.

Figure 81–8 illustrates (a) the increasing levels of estrogen secretion at puberty, (b) the cyclic variation during the monthly sexual cycles, (c) the further increase in estrogen secretion during the first few years of sexual life, (d) then progressive decrease in estrogen secretion toward the end of sexual life, and (e) finally almost no estrogen secretion beyond the menopause.

THE MENOPAUSE

At the age of 45 to 50 years the sexual cycles usually become irregular, and ovulation fails to occur during many of the cycles. After a few months to a few years, the cycles cease altogether, as illustrated in Figure 81–8. This period during which the cycles cease and the female sex hormones diminish rapidly to almost none at all is called the *menopause*.

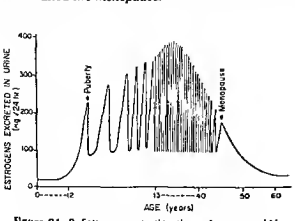


Figure 81–8. Estrogen secretion throughout sexual life.

The cause of the menopause is "burning out" of the ovaries. Throughout a woman's sexual life about 450 of the primordial follicles grow into vesicular follicles and ovulate, while literally thousands of the ova degenerate. At the age of about 45 years only a few primordial follicles still remain to be stimulated by FSH and LH, and the production of estrogens by the ovary decreases as the number of primordial follicles approaches zero (also illustrated in Figure 81–8). When estrogen production falls below a critical value, the estrogens can no longer inhibit the production of FSH and LH; nor can they cause an ovulatory surge of LH and FSH to cause oscillatory cycles. Instead, as illustrated in Figure 81–7, FSH and LH (mainly FSH) are produced thereafter *in large and continuous quantities*. Estrogens are produced in subcritical quantities for a short time after the menopause, but over a few years, as the final remaining primordial follicles become atretic, the production of estrogens by the ovaries falls almost to zero.

At the time of the menopause a woman must readjust her life from one that has been physiologically stimulated by estrogen and progesterone production to one devoid of these hormones The loss of the estrogens often causes marked physiological changes in the function of the body, including (1) "hot flashes" characterized by extreme flushing of the skin, (2) psychic sensations of dyspnea, (3) irritability, (4) fatigue, (5) anxiety, and (6) occasionally various psychotic states. These symptoms are of sufficient magnitude in approximately 15 per cent of women to warrant treatment If psychotherapy fails, daily administration of an estrogen in small quantities will reverse the symptoms, and, by gradually decreasing the dose, the postmenopausal woman is likely to avoid severe symptoms; unfortunately, such treatment can prolong the symptoms.

ABNORMALITIES OF SECRETION BY THE OVARIES

Hypogonadism. Less than normal secretion by the ovaries can result from poorly formed ovaries, lack of *ovaries, or genetically abnormal ovaries that secrete* the wrong hormones because of missing enzymes in the secretory cells. When ovaries are absent from birth or when they become nonfunctional before puberty, *female eunuchism* occurs. In this condition the usual secondary *sexual characteristics do not appear*, and the *sexual* organs remain infantile. Especially characteristic of this condition is prolonged growth of the long bones because the epiphyses do not unite with the shafts of these bones at as early an age as in the normal adolescent woman Consequently, the female eunuch is *essentially as tall* as, or perhaps even slightly taller than, her male counterpart of similar genetic background.

When the ovaries of a fully developed woman are removed, the sexual organs regress to some extent so that the uterus becomes almost *infantile in size, the* vagina becomes smaller, and the vaginal epithelium becomes thin and easily damaged. The breasts atrophy and become pendulous, and the pubic hair becomes thinner. These same changes occur in women after menopause.

Irregularity of Menses and Amenorrhea Due to Hypogonadism. As pointed out in the preceding discussion of the menopause, the quantity of estrogens

produced by the ovaries must rise above a critical value if they are to be able to cause rhythmical sexual cycles. Consequently, in hypogonadism or when the gonads are secreting small quantities of estrogens as a result of other factors, such as *hypothyroidism*, the ovarian cycle likely will not occur normally. Instead, several months may elapse between menstrual periods, or menstruation may cease altogether (amenorrhea). Characteristically, prolonged ovarian cycles are frequently associated with failure of ovulation, presumably due to insufficient secretion of luteinizing hormone, which is necessary for ovulation

Hypersecretion by the Ovaries. Extreme hypersecretion of ovarian hormones by the ovaries is a rare clinical entity, for excessive secretion of estrogens automatically decreases the production of gonadotropins by the pituitary, and this in turn limits the production of the ovarian hormones. Consequently, hypersecretion of feminizing hormones is recognized clinically only when a feminizing tumor develops

A rare granulosa cell tumor occasionally develops in an ovary, occurring more often after menopause than before. These tumors secrete large quantities of estrogens, which exert the usual estrogenic effects, including hypertrophy of the uterine endometrium and irregular bleeding from this endometrium. In fact, bleeding is often the first indication that such a tumor exists

THE FEMALE SEXUAL ACT

Stimulation of the Female Sexual Act. As is true in the male sexual act, successful performance of the female sexual act depends on both psychic stimulation and local sexual stimulation

As is also true in the male, the thinking of erotic thoughts can lead to female sexual desire, and this aids greatly in the performance of the female sexual act Such desire is probably based as much on one's background training as on physiological drive, though sexual desire does increase in proportion to the level of secretion of the sex hormones. Desire also changes during the sexual month, reaching a peak near the time of ovulation, probably because of the high levels of estrogen secretion during the preovulatory period

Local sexual stimulation in women occurs in more or less the same manner as in men, for massage, irritation, or other types of stimulation of the perineal region, sexual organs, and urinary tract create sexual sensations. The *clitoris* is especially sensitive for initiating sexual sensations. As in the male, the sexual sensory signals are mediated to the sacral segments of the spinal cord through the pudendal nerve and sacral plexus. Once these signals have entered the spinal cord, they are transmitted thence to the cerebrum. Also, local reflexes integrated in the sacral and lumbar spinal cord are at least partly responsible for female sexual reactions.

Female Erection and Lubrication. Located around the introitus and extending into the clitoris " tissue almost identical with the erectile tissue of the penis. This erectile tissue, like that of the penis, is controlled by the parasympathetic nerves that pass through the nervi erigentes from the sacral plexus to the external genitalia. In the early phases of sexual stimulation, parasympathetic signals dilate the arteries of the erectile tissues, and this allows rapid accumulation of blood in the erectile tissue so that the introitus tightens around the penis; this aids the male greatly in his attainment of sufficient sexual stimulation for ejaculation to occur.

Parasympathetic signals also pass to the bilateral Bartholin's glands located beneath the labia minora to cause secretion of mucus immediately inside the introitus This mucus is responsible for much of the lubrication during sexual intercourse, though much is also provided by mucus secreted by the vaginal epithelium as well and a small amount from the male urethral glands. The lubrication in turn is necessary for establishing during intercourse a satisfactory massaging sensation rather than an irritative sensation, which may be provoked by a dry vagina. A massaging sensation constitutes the optimal type of sensation for evoking the appropriate reflexes that culminate in both the male and female climaxes.

The Female Orgasm. When local sexual stimulation reaches maximum intensity, and especially when the local sensations are supported by appropriate psychic conditioning signals from the cerebrum, reflexes are initiated that cause the female orgasm, also called the *female climax*. The female orgasm is analogous to emission and ejaculation in the male, and it perhaps helps promota fertilization of the ovum Indeed, the human female is known to be somewhat more fertile when inseminated by normal sexual intercourse rather than by artificial methods, thus indicating an important function of the female orgasm. Possible effects that could result in this are:

First, during the orgasm the perineal muscles of the female contract rhythmically, which results from spinal cord reflexes similar to those that cause ejaculation in the male It is possible, also, that these same reflexes increase uterine and fallopian tube motility during the orgasm, thus helping transport the sperm toward the ovum, but the information on this subject is scanty. Also, the orgasm seems to cause dilation of the cervical canal for up to half an hour, thus allowing easy transport of the sperm

Second, in many lower animals, copulation causes the posterior pituitary gland to secrete oxytocin; this effect is probably mediated through the amygdaloid nuclei and then through the hypothalamus to the pituitary. The oxytocin in turn causes increased rhythmical contractions of the uterus, which has been postulated to cause rapid transport of the sperm Sperm have been shown to traverse the entire length of the fallopian tube in the row in approximately five minutes, a rate at

east ten times as fast as that which the swimming notions of the sperm themselves could achieve. Whether or not this occurs in the human female s unknown.

In addition to the possible effects of the orgasm n fertilization, the intense sexual sensations that levelop during the orgasm also pass to the cerebrum and cause intense muscle tension throughout he body. But after culmination of the sexual act, his gives way during the succeeding minutes to a ense of satisfaction characterized by relaxed peacefulness, an effect called *resolution*.

FEMALE FERTILITY

The Fertile Period of Each Sexual Cycle. The ovum remains viable and capable of being fertilized after it is expelled from the ovary probably no longer than 24 hours. Therefore, sperm must be available soon after ovulation if fertilization is to take place On the other hand, a few sperm can remain fertile in the female reproductive tract for up to 72 hours, though most of them for not more than 24 hours Therefore, for fertilization to take place, intercourse usually must occur some time between one day prior to ovulation up to one day after ovulation. Thus, the period of female fertility during each sexual cycle is short

The Rhythm Method of Contraception. One of the often practiced methods of contraception is to avoid intercourse near the time of ovulation The difficulty with this method of contraception is the impossibility of predicting the exact time of ovulation Yet the interval from ovulation until the next succeeding onset of menstruation is almost always between 13 and 15 days. Therefore, if the menstrual cycle is regular, with a periodicity of 28 days, ovulation usually occurs within one day of the 14th day of the cycle. If, on the other hand, the periodicity of the cycle is 40 days, ovulation usually occurs within one day of the 26th day of the cycle. Finally, if the periodicity of the cycle is 21 days, ovulation usually occurs within one day of the 7th day of the cycle. Therefore, it is usually stated that avoidance of intercourse for 4 days prior to the calculated day of ovulation and 3 days afterward prevents conception But such a method of contraception can be used only when the periodicity of the menstrual cycle is regular

Hormonal Suppression of Fertility—"The Pill." It has long been known that administration of either estrogen or progesterone, if given in appropriate quantity during the first half of the monthly female cycle, can inhibit ovulation. The reason for this is that administration of either of these can prevent the preovulatory surge of LH secretion by the pituitary gland, which, it will be recalled, is essential in causing ovulation.

The reason the administration of estrogen or progestenne prevents the preovulatory surge of LH secretion is not fully understood. However, experimental work has suggested that immediately before the surge occurs there is probably a sudden depression of estrogen secretion by the ovarian follicles, and that this might be the necessary signal for causing the positive feedback effect that leads to the surge. Obviously, administration of the sex hormones could prevent the initial hormonal depression that might be the initiating signal for ovulation.

The problem in devising methods for hormonal suppression of ovulation has been to develop appropriate combinations of estrogens and progestins that will suppress ovulation but that will not cause unwanted effects of these two hormones For instance, too much of either of the hormones can cause abnormal menstrual bleeding patterns. However, use of certain synthetic progestins in place of progesterone, especially the 19-norsteroids, along with small amounts of estrogens will usually prevent ovulation and yet also allow almost a normal pattern of menstruation. Therefore, almost all "pills" used for control of fertility consist of some combination of synthetic estrogens and synthetic progestins. The main reason for using synthetic estrogens and synthetic progestins is that the *natural* hormones are almost entirely destroyed by the liver within a short time after they are absorbed from the gastrointestinal tract into the portal circulation. However, many of the *synthetic* hormones can resist this destructive propensity of the liver, thus allowing oral administration.

Two of the most commonly used synthetic estrogens are *ethynyl estradiol* and *mestranol*. Among the most commonly used progestins are *norethindrone*, *norethynodrel*, *ethynodiol*, and *norgestrel* The medication is usually begun in the early stages of the monthly cycle and continued beyond the time that ovulation normally would have occurred. Then the medication is stopped, allowing menstruation to occur and a new cycle to begin.

Abnormal Conditions Causing Female Sterility. Approximately 1 of every 6 to 10 marriages is infertile; in about 60 per cent of these, the infertility is due to female sterility.

Occasionally, no abnormality whatever can be discovered in the female genital organs, in which case it must be assumed that the infertility is due either to abnormal physiological function of the genital system or to abnormal genetic development of the ova themselves.

However, probably by far the most common cause of female sterility is failure to ovulate This can result from either hyposecretion of gonadotropic hormones, in which case the intensity of the hormonal stimuli simply is not sufficient to cause ovulation, or it can result from abnormal ovaries that will not allow ovulation. For instance, thick capsules occasionally exist on the outside of the ovaries that prevent ovulation.

Because of the high incidence of anovulation in sterile women, special methods are often utilized to determine whether or not ovulation occurs These are all based on the effects of progesterone on the body, for the normal increase in progesterone secretion does not occur during the latter half of anovulatory cycles. In the absence of progesteronic effects, the cycle can be assumed to be anovulatory. One of these tests is simply to analyze the urine for a surge in pregnanediol, the end-product of progesterone metabolism, during the latter half of the sexual cycle, the lack of which indicates failure of ovulation. However, another common test is for the woman to chart her body temperature throughout the cycle. Secretion of progesterone during the latter half of the cycle raises the body temperature about 0.5°F, the temperature rise coming abruptly at the time of ovulation. Such a temperature chart, showing the point of ovulation, is illustrated in Figure 81–9.

Lack of ovulation caused by hyposecretion of the pituitary gonadotropic hormones can sometimes be treated by administration of *human chorionic gonadotropin*, a hormone that will be discussed in the following chapter and that is extracted from the human placenta.

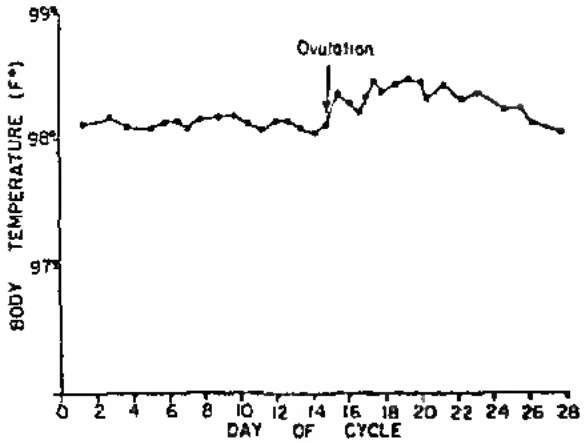


Figure 81–9. Elevation in body temperature shortly after ovulation.

This hormone, though secreted by the placenta, has almost exactly the same effects as luteinizing hormone and, therefore, is a powerful stimulator of ovulation. However, excess use of this hormone can cause ovulation from many follicles simultaneously, and this results in multiple births, an effect that has caused as many as seven children to be born to mothers treated for infertility with this hormone.

One of the most common causes of female sterility is *endometriosis*, a very common condition in which endometrial tissue almost identical to that of the uterine endometrium grows and even menstruates in the pelvic cavity surrounding the uterus, fallopian tubes, and ovaries. Endometriosis causes fibrosis throughout the pelvis, and this fibrosis sometimes so enshrouds the ovaries that an ovum cannot be released into the abdominal cavity. Often, also, endometriosis occludes the fallopian tubes, either at the fimbriated ends or elsewhere along their extent. Another common cause of female infertility is salpingitis, that is, inflammation of the fallopian tubes; this causes fibrosis in the tubes, thereby occluding them. In past years, such inflammation was extremely common as a result of gonococcal infection, but with modern therapy this is becoming a less prevalent cause of female infertility.

Finally, still another cause of infertility is abnormal secretion of mucus by the uterine cervix. Ordinarily, at the time of ovulation, the hormonal environment of estrogen causes secretion of a very thin mucus with special characteristics that will allow rapid mobility of sperm into the uterus and will actually guide the sperm up along mucous "threads." Abnormalities of the cervix itself, such as low-grade infection or inflammation, or abnormal hormonal stimulation of the cervix can lead to a viscous mucous plug that will prevent fertilization.

REFERENCES

Arnold, A. P., and Gorski, R. A. Gonadal steroid induction of structural sex differences in the central nervous system. *Annu. Rev. Neurosci.*, 7:413, 1984.

Beato, M. (ed.): Steroid Induced Uterine Proteins. New York, Elsevier/North Holland, 1980.

Behrman, H. R. Prostaglandins in hypothalamo-pituitary and ovarian function. *Annu. Rev. Physiol.*, 41:685, 1979.

Benirschke, K. The endometrium. *In* Yen, S. S. C., and Jaffe, R. B. (eds.): Reproductive Endocrinology. Philadelphia, W. B. Saunders Co., 1978, p. 241.

Chambon, P., *et al.* Promoter elements of genes coding for proteins and modulation of transcription by estrogens and progesterone. *Recent Prog. Horm. Res.*, 40:1, 1984.

Channing, C. P., *et al.* Ovarian follicular and luteal physiology. *In* Greep, R. O. (ed.): International Review of Physiology, Reproductive Physiology III. Vol. 22, Baltimore, University Park Press, 1980, p. 117.

Clark, J. H., and Peck, E. J., Jr.: Female Sex Steroids: Receptors and Function. New York, Springer-Verlag, 1979.

Conn, P. M., *et al.* Gonadotropin releasing hormone: Molecular and cell biology, physiology, and clinical applications. *Fed. Proc.*, 43:2351, 1984.

Coutts, S. R. (ed.): Functional Morphology of the Human Ovary. Baltimore, University Park Press, 1981.

Cummings, A. M., and Yochim, J. M. Differentiation of the uterus in preparation for gestation: A model for the action of progesterone. *J. Theor. Biol.*, 106:353, 1984.

Eaweett, D. W., and Bedford, J. M. (eds.): The Spermatozoon. Baltimore, Urban & Schwarzenberg, 1979.

Fishman, J. Aromatic hydroxylation of estrogens. *Annu. Rev. Physiol.*, 45:61, 1983.

Glass, R. H. Infertility. *In* Yen, S. S. C., and Jaffe, R. B. (eds.): Reproductive Endocrinology. Philadelphia, W. B. Saunders Co., 1978, p. 398.

Greenblatt, R. B. (ed.): Induction of Ovulation. Philadelphia, Lea & Febiger, 1979.

Kase, N. G., and Speroff, L. The ovary. *In* Bondy, P. K., and Rosenberg, L. E. (eds.): Metabolic Control and Disease, 8th Ed. Philadelphia, W. B. Saunders Co., 1980, p. 1579.

Krieger, D. T. (ed.): Current Therapy in Endocrinology, 1983–1984. St. Louis, C. V. Mosby, 1984.

Leon, A. The Cycling Female: Her Menstrual Rhythm. San Francisco, W. H. Freeman, 1979.

Leung, P. C. K., and Armstrong, D. T. Interactions of steroids and gonadotropins in the control of steroidogenesis in the ovarian follicle. *Annu. Rev. Physiol.*, 42:71, 1980.

McCann, S. M. Physiology and pharmacology of LHRH and somatostatin. *Annu. Rev. Pharmacol. Toxicol.*, 22:491, 1982.

McEwen, B. S. Binding and metabolism of sex steroids by the hypothalamic pituitary unit: Physiological implications. *Annu. Rev. Physiol.*, 42:97, 1980.

McEwen, B. S., and Parsons, B. Gonadal steroid action on the brain: Neurochemistry and neuropharmacology. *Annu. Rev. Pharmacol. Toxicol.*, 22:555, 1982.

Meschia, G. Circulation to female reproductive organs. *In* Shepherd, J. T., and Abboud, F. M. (eds.): Handbook of Physiology. Sec. 2, Vol. III. Bethesda, American Physiological Society, 1983, p. 241.

Midgley, A. R., and Sadler, W. A. (eds.): Ovarian Follicular Development and Function. New York, Raven Press, 1979.

Mishell, D. R., Jr. (ed.): Long Acting Steroid Contraception. New York, Raven Press, 1983.

Mishell, D. R., Jr., and Davajan, V. (eds.): Reproductive Endocrinology, Infertility, and Contraception. Philadelphia, F. A. Davis Co., 1979.

Molinatti, G. M. (ed.): Androgenization in Women: Pathophysiology and Clinical Aspects. New York, Raven Press, 1983.

Ochiai, K. (ed.): Endocrine Correlates of Reproduction. New York, Springer-Verlag, 1984.

Odell, W. D.: FSH. *In* DeGroot, L. J., *et al.* (eds.): Endocrinology. Vol. 1. New York, Grune & Stratton, 1979, p. 149.

Odell, W. D.: LH. *In* DeGroot, L. J., *et al.* (eds.): Endocrinology. Vol. 1. New York, Grune & Stratton, 1979, p. 151.

Pohl, C. R., and Knobil, E. The role of the central nervous system in the control of ovarian function in higher primates. *Annu. Rev. Physiol.*, 44:583, 1982.

Reiter, E. O., and Grumbach, M. M. Neuroendocrine control mechanisms and the onset of puberty. *Annu. Rev. Physiol.*, 44:595, 1982.

Richards, J. S. Maturation of ovarian follicles: Actions and interactions of pituitary and ovarian hormones on follicular cell differentiation. *Physiol. Rev.*, 60:51, 1980.

Sanborn, B. M., *et al.* The role of the cyclic nucleotides in reproductive processes. *Annu. Rev. Physiol.*, 42:37, 1980.

Schwartz, N. B. (ed.): Dynamics of Ovarian Function. New York, Raven Press, 1981.

Serra, G. B. (ed.): The Ovary. New York, Raven Press, 1983.

Tollison, C. D., and Adams, H. E. Sexual Disorders: Treatment, Theory and Research. New York, Gardner Press, 1979.

Weiss, G.: Relaxin. *Annu. Rev. Physiol.*, 46:43, 1984.

82

Pregnancy and Lactation

In the preceding two chapters the sexual functions of the male and female were described to the point of fertilization of the ovum. If the ovum becomes fertilized, a completely new sequence of events called *gestation*, or *pregnancy*, takes place, and the fertilized ovum eventually develops into a full-term fetus. The purpose of the present chapter is to discuss the early stages of ovum development *after* fertilization and then to discuss the physiology of pregnancy. In the following chapter some special problems of fetal and early childhood physiology are discussed.

MATURATION OF THE OVUM

Shortly before the *primary oocyte* (the original stage of the ovum in the ovary) is released from the follicle, its nucleus divides by meiosis and a so-called *first polar body* is expelled from the nucleus of the oocyte, and this then becomes the *secondary oocyte*. In this process, each of the 23 pairs of chromosomes loses one of the partners to the polar body so that 23 *unpaired* chromosomes remain in the secondary oocyte. A few hours after the sperm enters the oocyte the nucleus divides again, and a *second polar body* is expelled, thus forming the *mature ovum*. This is also a meiotic division so that there still remain 23 unpaired chromosomes.

One of the 23 chromosomes is always a female chromosome, called an *X chromosome*. When this combines with a sperm that carries an X chromosome, giving an XX combination, a female child is formed, as was explained in Chapter 80. And, when the X chromosome of the ovum is paired with a sperm that carries a Y chromosome, giving an XY combination, a male child is formed.

TRANSPORT, FERTILIZATION, AND IMPLANTATION OF THE DEVELOPING OVUM

Entry of the Ovum into the Fallopian Tube. When ovulation occurs, the ovum, along with its attached granulosa cells, the *corona radiata*, is expelled directly into the peritoneal cavity and must then enter one of the fallopian tubes. The fimbriated ends of each fallopian tube fall naturally around the ovaries; the inner surfaces of the fimbriated tentacles are lined with ciliated epithelium, and the *cilia* continually beat toward the *abdominal ostium* of the fallopian tube. One can actually see a slow fluid current flowing toward the ostium. By this means the ovum enters one or the other fallopian tube.

It would seem likely that many ova might fail to enter the fallopian tubes. However, on the basis of conception studies it is probable that as many as 98 to 100 per cent succeed in this task. Indeed, cases are on record in which women with one ovary removed and the opposite fallopian tube removed have had as many as four children with relative ease of conception, thus illustrating that ova can even enter the opposite fallopian tube.

Fertilization of the Ovum. After ejaculation, at least some sperm are transported through the uterus to the ovarian end of the fallopian tubes within 5 to 10 minutes, probably aided by contractions of the uterus and fallopian tubes during and following the female orgasm. However, of the half billion sperm deposited in the vagina only 1000 to 3000 succeed in actually reaching the proximity of the ovum.

Fertilization of the ovum normally takes place soon after the ovum enters the fallopian tube. However, before a sperm can enter the ovum, it must first penetrate the multiple layers of granulosa cells attached to the outside of the ovum, called the *corona radiata*. This is believed to require the hyaluronidase and proteolytic enzymes released by the acrosomes of the sperm to open a pathway through which at least one sperm can pass. In fact, even though only one sperm is required to fertilize the ovum, if an insufficient number of sperm are released in the semen the male donor of the semen is likely to be infertile. This is believed to result from the fact that many sperm are required to attack the corona radiata before one can succeed in penetrating the barrier. In addition to the enzymes from the sperm, bicarbonate ions in the fallopian tube secretions probably also help in dispersing the granulosa cells.

Almost never does more than one sperm enter the ovum for the following reason: The zona r

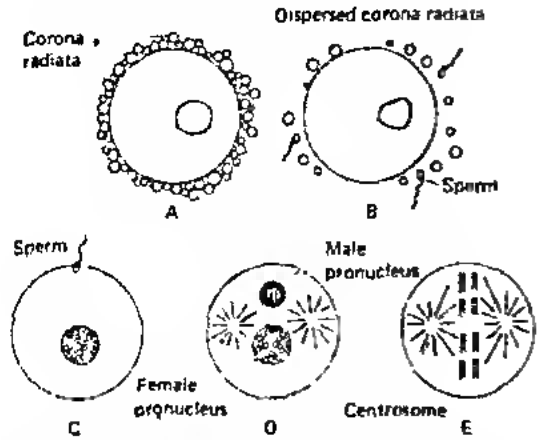


Figure 82–1. Fertilization of the ovum, showing *A*, the mature ovum surrounded by the corona radiata, *B*, dispersal of the corona radiata, *C* entry of the sperm, *D* formation of the male and female pronuclei and *E*, reorganization of a full complement of chromosomes and beginning division of the ovum. (Modified from Arey: Developmental Anatomy, 7th ed. Philadelphia, W. B. Saunders Company, 1974.)

lucida of the ovum, the layer of mucopolysaccharide surrounding the ovum, has a lattice-type structure, and once this is punctured, some substance (perhaps one of the proteolytic enzymes of the sperm acrosome) seems to diffuse into the lattice to prevent penetration by additional sperm. Indeed, microscopic studies show that many additional sperm do attempt to penetrate the zona pellucida but become inactivated after traveling only part way through.

Once a sperm enters the ovum, its head swells rapidly to form a *male pronucleus*, which is illustrated in Figure 82–1. Later, after the nucleus of the ovum has discharged the second polar body, the 23 chromosomes of the male pronucleus and the 23 of the *female pronucleus* align themselves to reform a complete complement of 46 chromosomes (23 pairs) in the fertilized ovum.

Transport in the Fallopian Tubes

After fertilization has occurred, an additional three to five days are normally required for transport of the ovum through the fallopian tube into the cavity of the uterus. This transport is effected mainly by a feeble fluid current in the tube resulting from action of the ciliated epithelium that lines the tube, the cilia always beating toward the uterus. It is possible also that weak contractions of the fallopian tube aid in the passage of the ovum.

The fallopian tubes are lined with a rugged, cryptoid surface that actually impedes the passage of the ovum despite the fluid current. Also, the isthmus of the fallopian tube (the last 2 centimeters before the uterus is entered) remains spastically contracted for the first three days following ovulation. After this time, probably under the relaxing influence of the rapidly increasing progesterone, this region loses the spasm, thus allowing entry of the ovum to the uterus.

The delayed transport of the ovum through the fallopian tube allows several stages of division to occur before the ovum, now called a *zygote*, enters the uterus. During this time, large quantities of secretions are formed by secretory cells that alternate with ciliated cells lining the fallopian tube. These secretions are for nutrition of the zygote. Indeed, a zygote will continue to divide in vitro as long as it is bathed in a solution of homogenized fallopian tube mucosa.

Implantation of the Zygote in the Uterus

After reaching the uterus, the developing zygote usually remains in the uterine cavity an additional two to four days before it implants in the endometrium; thus, implantation ordinarily occurs about the seventh day following ovulation. Before implantation, the zygote obtains its nutrition from the endometrial secretions called "uterine milk" and it grows to be a solid mass of cells, a hundred or more, called the *morula*. It is at this stage that implantation occurs.

Implantation results from the action of *trophoblast cells* that develop over the surface of the morula. These cells secrete proteolytic enzymes that digest and liquefy the cells of the endometrium. The fluid and nutrients thus released are then actively transported by the same trophoblastic cells into the morula, adding still further sustenance for growth.

Shortly after implantation, a cavity develops in the mass of cells, and the embryo begins to develop along one wall of the cavity. This is the *blastocyst stage*. Figure 82–2 shows an early implanted human blastocyst, with a very small embryo illustrated in color.

Once implantation has taken place, the trophoblast and other sublying cells proliferate rapidly

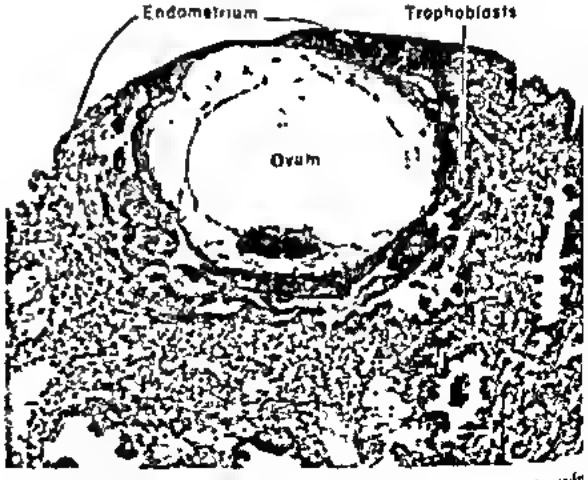


Figure 82–2. Implantation of the early human embryo, showing trophoblastic digestion and invasion of the endometrium. (Courtesy of Dr. Arthur Hertig.)

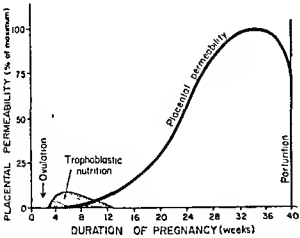


Figure 82–3. Nutrition of the fetus, illustrating that most of the early nutrition is due to trophoblastic digestion and absorption of ... from the endometrial decidua and that essentially all the later nutrition results from diffusion through the placental membrane.

...rming the placenta and the various membranes ... pregnancy.

EARLY INTRAUTERINE NUTRITION OF THE EMBRYO

In the previous chapter it was pointed out that the progesterone secreted during the latter half of each sexual cycle has a special effect on the endometrium to convert the endometrial stromal cells into large swollen cells that contain extra quantities of glycogen, proteins, lipids, and even some necessary minerals for development of the zygote. If a zygote implants in the endometrium, the continued secretion of progesterone causes the stromal cells to swell still more and to store even more nutrients. These cells are now called *decidual cells*, and the total mass of cells is called the *decidua*.

As the trophoblast cells invade the decidua, digesting and imbibing it, the stored nutrients in the decidua are used by the embryo for appropriate growth and development. During the first week after implantation, this is the only means by which the embryo can obtain nutrients, and the embryo continues to obtain a large measure of its total nutrition in this way for 8 to 12 weeks, though the placenta also begins to provide slight amounts of nutrition after approximately the 16th day beyond fertilization (a little over a week after implantation). Figure 82–3 depicts this trophoblastic period of nutrition, which gradually gives way to placental nutrition.

FUNCTION OF THE PLACENTA

DEVELOPMENTAL AND PHYSIOLOGIC ANATOMY OF THE PLACENTA

While the trophoblastic cords from the blastocyst are attaching to the uterus, blood capillaries grow into the cords from the vascular system of the embryo, and, by the 16th day after fertilization, blood begins to flow. Simultaneously, *blood sinuses* supplied with blood from the mother develop around the trophoblastic cords. The trophoblast cells then gradually send out more and more projections, which become the *placental villi* into which fetal capillaries grow. Thus, the villi, carrying fetal blood, are surrounded by sinuses containing maternal blood.

The final structure of the placenta is illustrated in Figure 82–4. Note that the fetus' blood flows through two *umbilical arteries*, then to the capillaries of the villi, and thence back through the *umbilical vein* into the fetus. At the same time, the mother's blood flows from the *uterine arteries* into large *maternal sinuses* surrounding the villi and then back into the *uterine veins* of the mother.

The lower part of Figure 82–4 illustrates the relationship between the fetal blood of the villus and the blood of the mother in the fully developed placenta. The capillaries of the villus are lined with an extremely thin endothelium and are surrounded by a layer of *mesenchymal tissue* that is covered on the outside of the villus by a layer of *syncytial trophoblast* cells. During the first 16 weeks of pregnancy, still an additional layer of cells is present immediately beneath the syncytial trophoblast layer. This layer is composed of distinct cuboidal cells called *cytotrophoblast cells*, or *Langerhans' cells*.

The total surface area of all the villi of the mature

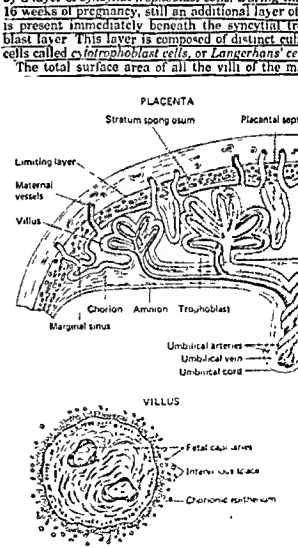


Figure 82–4. Above: Organization of the mature placenta. Below: Relationship of the fetal blood in the villus capillaries to the mother's blood in the intervillous spaces. (Modified from Gray and Goss: Anatomy of the Human Body. Lea & Febiger and from Arey: Developmental Anatomy. 7th ed. W. B. Saunders.)

placenta is only a few square meters—many times less than the area of the pulmonary membrane. Also, remember that even at full maturity the placental membrane is still several cell layers thick, and the minimum distance between the maternal blood and the fetal blood is 3 5 microns, almost ten times the distance across the alveolar membranes of the lung. Nevertheless, many nutrients and other substances pass through the placental membrane by diffusion in very much the same manner that diffusion occurs through the alveolar membranes of the lungs and the capillary membranes elsewhere in the body

PERMEABILITY OF THE PLACENTAL MEMBRANE

The major function of the placenta is to allow *diffusion* of foodstuffs from the mother's blood into the fetus' blood and diffusion of excretory products from the fetus back into the mother. Therefore, it is important to know the degree of permeability of the placental membrane. *Permeability is expressed as the total quantity of a given substance that crosses the entire placental membrane in a given time for a given concentration difference.*

In the early months of development, placental permeability is relatively slight, as illustrated in Figure 82–3, for two reasons· First, the total surface area of the placental membrane is still small at that time, and, second, the villar membranes have not yet been reduced to their minimum thickness However, as the placenta becomes older, the permeability increases progressively until the last month or so of pregnancy when it begins to decrease again This decrease results from deterioration of the placenta caused by its age and sometimes from destruction of whole segments due to infarction.

Rarely, "breaks" occur in the placental membrane, which allow fetal blood cells to pass into the mother or, more rarely, the mother's cells to pass into the fetus Indeed, there are instances in which the fetus bleeds severely into the mother's circulation because of a ruptured placental membrane.

Diffusion of Oxygen Through the Placental Membrane. Almost exactly the same principles are applicable for the diffusion of oxygen through the placental membrane as through the pulmonary membrane; these principles were discussed in Chapter 40. The dissolved oxygen in the blood of the large maternal sinuses simply passes into the fetal blood because of an oxygen pressure gradient from the mother's blood to the fetus' blood The mean P_{O_2} in the mother's blood in the maternal sinuses is approximately 50 mm Hg toward the end of pregnancy, and the mean P_{O_2} in the fetal blood after it becomes "oxygenated" is about 30 mm Hg Therefore, the mean pressure gradient for diffusion of oxygen through the placental membrane is about 20 mm Hg.

One might wonder how it is possible for a fetus to obtain sufficient oxygen when the fetal bloo leaving the placenta has a P_{O_2} of only 30 mm Hg However, there are three different reasons wh even this low P_{O_2} is capable of allowing the fet blood to transport almost as much oxygen to th fetal tissues as is transported by the mother's bloo to her tissues:

First, the hemoglobin of the fetus is mainly *fet hemoglobin,* a type of hemoglobin synthesized i the fetus prior to birth. Figure 82–5 illustrates th comparative oxygen dissociation curves of mate nal hemoglobin and fetal hemoglobin, showin that the curve for fetal hemoglobin is shifted t the left of that for maternal hemoglobin. T means that at a given P_{O_2}, the fetal hemoglob can carry as much as 20 to 30 per cent mor oxygen than can maternal hemoglobin.

Second, the *hemoglobin concentration of the fetu is about 50 per cent greater than that of the mothe* therefore, this is an even more important facto than the first in enhancing the amount of oxyge transported to the fetal tissues.

Third, the *Bohr effect,* which was explained i relation to the exchange of carbon dioxide an oxygen in the lung in Chapter 41, provides anothe factor that enhances the transport of oxygen b the fetal blood. That is, hemoglobin can carry mor oxygen at a low P_{CO_2} than it can at a high P_{CO} The fetal blood entering the placenta carries larg amounts of carbon dioxide, but much of this carbo dioxide diffuses from the fetal blood into the ma ternal blood Loss of the carbon dioxide makes th fetal blood more alkaline while the increased car bon dioxide in the maternal blood makes this mor acidic. These changes cause the combining capac ity of fetal blood for oxygen to become increase and that of the maternal blood to become de creased. This forces more oxygen from the mater nal blood while enhancing the oxygen in the feta blood. Thus, the Bohr shift operates in one direc tion in the maternal blood and in the other in th fetal blood, these two effects adding to make th Bohr shift twice as important here as it is fo

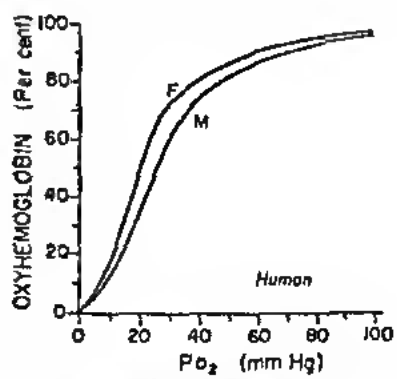


Figure 82–5. Oxygen-hemoglobin dissociation curves for maternal *(M)* and fetal *(F)* bloods, illustrating the ability of the feta blood to carry a much greater quantity of oxygen than car maternal blood for a given blood P_{O_2} (From Metcalfe, Moll and Bartels *Fed. Proc.*, 23 775, 1964)

oxygen exchange in the lungs; therefore, it is actually called the *double Bohr effect*.

By these three means, the fetus is capable of receiving more than adequate oxygen through the placenta despite the fact that the fetal blood leaving the placenta has a PO_2 of only 30 mm Hg.

The total diffusing capacity of the placenta for oxygen at term is about 1.2 ml of oxygen per minute per millimeter of hemoglobin oxygen gradient. This compares favorably with that of the lungs of the newborn baby.

Diffusion of Carbon Dioxide Through the Placental Membrane. Carbon dioxide is continually formed in the tissues of the fetus in the same way that it is formed in maternal tissues. And the only means for excreting the carbon dioxide is through the placenta. The PCO_2 of the fetal blood is 2 to 3 mm Hg higher than that of the maternal blood. This small pressure gradient for carbon dioxide across the membrane is sufficient to allow adequate diffusion of carbon dioxide because the extreme solubility of carbon dioxide in the tissues allows carbon dioxide to diffuse about 20 times as rapidly as oxygen.

The PCO_2 of the blood in the maternal sinuses is usually less than the normal value of 40 mm Hg found in the arterial blood of the nonpregnant woman because the estrogens and progesterone of pregnancy overdrive the mother's breathing, thus causing her to blow off excess CO_2 from her lungs This obviously helps keep the PCO_2 in the fetus' blood at a low level aa well, usually very near to the normal value of 40 mm Hg.

Diffusion of Foodstuffs Through the Placental Membrane. Other metabolic substrates needed by the fetus diffuse into the fetal blood in the same manner as oxygen does. For instance, in the late stages of pregnancy the fetus often utilizes more glucose than does the entire body of the mother. To provide this much glucose, the trophoblast cells lining the placental villi provide for *facilitated diffusion* of glucose through the placental membrane. That is, the glucose is transported by carrier molecules in the trophoblast cell membrane. Even so, the glucose level in the fetal blood is still 20 to 30 per cent lower than in the maternal blood.

Because of the high solubility of fatty acids in cell membranes, these also diffuse from the maternal blood into the fetal blood but more slowly than glucose so that glucose is preferentially used by the fetus for nutrition. Also, such substances as ketone bodies and potassium, sodium, and chloride ions diffuse from the maternal blood into the fetal blood.

Active Absorption by the Placental Membrane. As pointed out previously, early nutrition of the embryo depends on phagocytosis of fallopian tube and uterine secretions, and even on phagocytosis of the endometrial decidua. The trophoblast cells that line the outer surface of the villi can also *actively absorb* certain nutrients from the maternal blood in the placenta, at least during the first half of pregnancy and perhaps even throughout the entire pregnancy. For instance, the measured *amino acid* content of fetal blood is greater than that of maternal blood, and *calcium* and *inorganic phosphate* occur in greater concentration in fetal blood than in maternal blood, while *ascorbic acid* is as much as three times as concentrated in the fetal blood. These findings indicate that the placental membrane has the ability to absorb actively at least small amounts of certain substances even during the latter part of pregnancy.

Excretion Through the Placental Membrane. In the same manner that carbon dioxide diffuses from the fetal blood into the maternal blood, other excretory products formed in the fetus diffuse in the opposite direction into the maternal blood and then are excreted along with the excretory products of the mother. These include especially the *nonprotein nitrogens* such as *urea, uric acid,* and *creatinine*. The level of urea in the fetal blood is only slightly greater than that in maternal blood because urea diffuses through the placental membrane with considerable ease. On the other hand, creatinine, which does not diffuse as easily, has a considerably higher concentration gradient percentagewise. Therefore, excretion from the fetus occurs mainly if not entirely as a result of diffusion gradients across the placental membrane—that is, because of higher concentrations of the excretory products in the fetal blood than in the maternal blood

HORMONAL FACTORS IN PREGNANCY

In pregnancy, the placenta forms large quantities of *human chorionic gonadotropin, estrogens, progesterone, and human chorionic somatomammotropin*, the first three of which, and perhaps the fourth as well, are essential to the continuance of pregnancy.

HUMAN CHORIONIC GONADOTROPIN AND ITS EFFECT IN CAUSING PERSISTENCE OF THE CORPUS LUTEUM AND IN PREVENTING MENSTRUATION

Menstruation normally occurs approximately 14 days after ovulation, at which time most of the endometrium of the uterus sloughs away from the uterine wall and is expelled to the exterior. If this should happen after an ovum has implanted, the pregnancy would terminate. However, this is prevented by the secretion of human chorionic gonadotropin in the following manner:

Coincidently with the development of the trophoblast cells from an early fertilized ovum, the hormone *human chorionic gonadotropin* is secreted by the syncytial trophoblast cells into the fluids of the mother. As illustrated in Figure 82–6, the secretion of this hormone can first be measured in the blood 8 to 9 days after ovulation, shortly after the zygote implants in the endometrium. Then the rate of secretion rises rapidly to reach a maximum approximately 7 to 9 weeks after ovulation, and

stretch, perineal stretch, and stretch or tearing of structures in the vaginal canal itself. This pain is conducted by somatic nerves instead of by the visceral sensory hypogastric nerves.

INVOLUTION OF THE UTERUS

During the first four to five weeks following parturition, the uterus involutes. Its weight becomes less than one half its immediate postpartum weight within a week, and in four weeks, if the mother lactates, the uterus may be as small as it had been prior to pregnancy. This effect of lactation results from the suppression of gonadotropin and ovarian hormone secretion during the first few months of lactation, as will be discussed later. During early involution of the uterus the placental site on the endometrial surface autolyzes, causing a vaginal discharge known as "lochia," which is first bloody and then serous in nature, continuing in all for approximately a week and a half. After this time, the endometrial surface will have become re-epithelialized and ready for normal, nongravid sex life again.

LACTATION

DEVELOPMENT OF THE BREASTS

The breasts begin to develop at puberty; this development is stimulated by the estrogens of the monthly sexual cycles that stimulate growth of the stroma and ductal system plus deposition of fat to give mass to the breasts. However, much additional growth occurs during pregnancy, and the glandular tissue only then becomes completely developed for actual production of milk.

Growth of the Ductal System—Role of the Estrogens. All through pregnancy, the tremendous quantities of estrogens secreted by the placenta cause the ductal system of the breasts to grow and to branch. Simultaneously, the stroma of the breasts also increases in quantity, and large quantities of fat are laid down in the stroma.

However, also important in growth of the ductal system are at least four other hormones: *growth hormone, prolactin,* the *adrenal glucocorticoids,* and *insulin.* Each of these is known to play at least some role in protein metabolism, which presumably explains their function in the development of the breasts.

Development of the Lobule-Alveolar System—Role of Progesterone. Final development of the breasts into milk-secreting organs also requires the additional action of *progesterone.* Once the ductal system has developed, progesterone, acting synergistically with all the other hormones just mentioned, causes growth of the lobules, budding of alveoli, and development of secretory characteristics in the cells of the alveoli. These changes are analogous to the secretory effects of progesterone on the endometrium of the uterus during the latter half of the female sexual cycle.

INITIATION OF LACTATION—FUNCTION OF PROLACTIN

Though estrogen and progesterone are essentia for the physical development of the breasts durin pregnancy, a specific effect of both these hormone is to inhibit *the actual secretion of milk.* On th other hand, the hormone *prolactin* has exactly th opposite effect, promotion of milk secretion. Thi hormone is secreted by the mother's pituitar gland, and its concentration in her blood rise steadily from the fifth week of pregnancy unt birth of the baby, at which time it has risen t very high levels, about ten times the normal non pregnant level. This very high level of prolacti at the end of pregnancy is illustrated in Figur 82–9. In addition, the placenta secretes large quan tities of *human chorionic somatomammotropi* which also has mild lactogenic properties, thu supporting the prolactin from the mother's pitu tary. Even so, because of the suppressive effects c estrogen and progesterone only a few milliliters c fluid are secreted each day until after the baby i born. This fluid is called *colostrum;* it contain essentially the same concentrations of proteins an lactose as milk but almost no fat, and its maximu rate of production is about 1/100 the subsequen rate of milk production.

Immediately after the baby is born, the sudde loss of both estrogen and progesterone secretion b the placenta now allows the lactogenic effect o the prolactin from the mother's pituitary gland t assume its natural milk-promoting role, an within two or three days the breasts begin t secrete copious quantities of milk instead of colos trum. This secretion of milk requires an adequat background secretion of most of the mother's othe hormones as well, but most important of all ar *growth hormone, cortisol,* and *parathyroid hor mone.* These hormones are necessary to provid

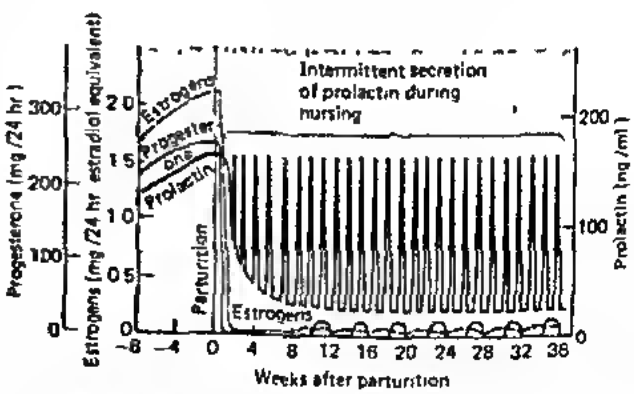


Figure 82–9. Changes in rates of secretion of estrogens, progesterone, and prolactin for 8 weeks prior to parturition and for 36 weeks thereafter. Note especially the decrease of prolactin secretion back to basal levels within a few weeks but also the intermittent periods of marked prolactin secretion (for about one hour at a time) during and after periods of nursing.

the amino acids, fatty acids, glucose, and calcium that are required for milk formation.

Following birth of the baby, the *basal level* of prolactin secretion returns during the next few weeks to the nonpregnant level, as shown in Figure 82–9. However, each time the mother nurses her baby, nervous signals from the nipples to the hypothalamus cause approximately a tenfold surge in prolactin secretion lasting about one hour, which is also shown in the figure. This prolactin in turn acts on the breasts to provide the milk for the next nursing period. If this prolactin surge is absent, if it is blocked as a result of hypothalamic or pituitary damage, or if nursing does not continue, the breasts lose their ability to produce milk within a few days. However, milk production can continue for several years if the child continues to suckle, though the rate of milk formation normally decreases considerably within seven to nine months.

Hypothalamic Control of Prolactin Secretion. The hypothalamus plays an essential role in controlling prolactin secretion as it does for the control of secretion of almost all the other anterior pituitary hormones as well. However, this control is different in one aspect: the hypothalamus mainly *stimulates* the production of all the other hormones, but it mainly *inhibits* prolactin production. Consequently, damage to the hypothalamus or blockage of the hypothalamic-hypophysial portal system increases prolactin secretion while it depresses secretion of the other anterior pituitary hormones. Yet, under special conditions, such as when the baby suckles the breast, it seems that a different type of signal from the hypothalamus can then increase the secretion of prolactin.

Therefore, it is believed that two different factors formed in the hypothalamus are transported to the anterior pituitary through the hypothalamic-hypophysial portal system to control prolactin release by the anterior pituitary gland. These are called *prolactin inhibitory factor* (PIF), which is the dominant hormone under most normal conditions, and *prolactin-releasing factor* (PRF), which can intermittently increase prolactin secretion. Though neither of these two hormones has been identified with certainty, PIF is believed to be *dopamine*, which is known to be secreted in the hypothalamus and which can decrease prolactin secretion as much as tenfold.

Suppression of the Female Sexual Cycle During Nursing. In about half of nursing mothers, the ovarian cycle and ovulation do not resume until a few weeks after cessation of nursing the baby. Though the exact cause of this is not known, it is likely that the same nervous signals from the breasts to the hypothalamus that cause prolactin secretion during suckling simultaneously inhibit secretion of luteinizing hormone releasing hormone by the hypothalamus, which in turn suppresses the formation of the pituitary gonadotropic hormones, luteinizing hormone and follicle-stimulating hormone. Yet, after several months of lactation, in about half of the mothers the pituitary begins again to secrete sufficient gonadotropic hormones to reinstate the monthly sexual cycle. The rhythmic interplay between the ovarian and pituitary hormones during the sexual month does not require marked reduction in prolactin secretion.

THE EJECTION OR "LET-DOWN" PROCESS IN MILK SECRETION—FUNCTION OF OXYTOCIN

Milk is secreted continuously into the alveoli of the breasts, but milk does not flow easily from the alveoli into the ductal system and therefore does not continually leak from the breast nipples. Instead, the milk must be ejected, or "let-down," from the alveoli to the ducts before the baby can obtain it. This process is caused by a combined neurogenic and hormonal reflex involving the posterior pituitary hormone *oxytocin*.

When the baby suckles the breast, sensory impulses are transmitted through somatic nerves from the nipples to the spinal cord and then to the hypothalamus, there causing *oxytocin* secretion at the same time that they cause prolactin secretion. The oxytocin is then carried in the blood to the breasts where it causes the *myoepithelial cells* that surround the outer walls of the alveoli to contract, thereby expressing the milk from the alveoli into the ducts. Thus, within 30 seconds to a minute after a baby begins to suckle the breast, milk begins to flow. This process is called milk ejection, or milk let-down.

Suckling on one breast causes milk flow not only in that breast but also in the opposite breast. It is especially interesting that fondling the baby by the mother or hearing the baby crying also often gives enough of a signal to her hypothalamus to cause milk ejection.

Inhibition of Milk Ejection. A particular problem in nursing the baby comes from the fact that many psychogenic factors or generalized sympathetic stimulation throughout the body can inhibit oxytocin secretion and consequently depress milk ejection. For this reason, the mother must have an undisturbed puerperium if she is to be successful in nursing her baby.

MILK AND THE METABOLIC DRAIN ON THE MOTHER CAUSED BY LACTATION

Table 82–1 lists the contents of human milk and cow's milk. The concentration of lactose in human milk is approximately 50 per cent greater than that in cow's milk, but on the other hand the concentration of protein in cow's milk is ordinarily two or more times as great as that in human milk

Table 82–1 PERCENTAGE COMPOSITION OF MILK

	Human Milk	Cow's Milk
Water	88.5	87
Fat	3.3	3.5
Lactose	6.8	4.8
Casein	0.9	2.7
Lactalbumin and other protein	0.4	0.7
Ash	0.2	0.7

Finally, the ash, which contains the calcium and other minerals, is only one third as much in human milk as in cow's milk.

At the height of lactation 1.5 liters of milk may be formed each day. With this degree of lactation great quantities of metabolic substrates are drained from the mother. For instance, approximately 50 grams of fat then enter the milk each day; and approximately 100 grams of lactose, which must be derived from glucose, are lost from the mother each day. Also, 2 to 3 grams of calcium phosphate may be lost each day, and, unless the mother is drinking large quantities of milk and has an adequate intake of vitamin D, the output of calcium and phosphate by the lactating mammaa will be much greater than the intake of these substances. To supply the needed calcium and phosphate, the parathyroid glands enlarge greatly, and the bones become progressively decalcified. The problem of decalcification is usually not very great during pregnancy, but it can be a distinct problem during lactation

REFERENCES

Angeh, A (ed) Endocrinology of Cystic Breast Disease New York, Raven Press 1983

Aragona, C, and Friesen, H G Lactation and galactorrhea *In* DeGroot L J, *et al* (eds) Endocrinology Vol 3 New York, Grune & Stratton 1979 p 1613

Aynsley-Green, A Metabolic and endocrine interrelations in the human fetus and neonate. *Am J Clin Nutr*, *41*:299, 1985

Battaglia, F C and Meschia, G Principal substrates of fetal metabolism *Physiol Rev* *58* 499, 1978

Beaconsfield P, *et al* (eds) Placenta. A Neglected Experimental Animal New York, Pergamon Press, 1979

Beckman, D A, and Brent, R L Mechanisms of teratogenesis. *Annu Rev Pharmacol Toxicol*, *24* 483, 1984

Buster, J E, and Marshall, J R Conception, gamete and ovum transport, implantation, fetal placental hormones, hormonal preparation of parturition and parturition control *In* DeGroot, L J, *et al* (eds.) Endocrinology Vol 3 New York, Grune & Stratton, 1979, p 1535

Cannon, J G Structure-activity relationships of dopamine agonists *Annu Rev Pharmacol Toxicol*, *23* 103, 1983

Challis, J R. G Endocrinology of late pregnancy and parturition *In* Greep, R. O (ed) International Review of Physiology Reproductive Physiology III Vol 22 Baltimore, University Park Press 1980, p 277

Chamberlain, G, and Wilkinson, A (eds) Placenta Transfer. Baltimo University Park Press, 1979

Clark, D A, *et al* Local active suppression by suppressor cells in t decidua A review *Am J Reprod Immunol*, *6* 78, 1984

Conn, P M, *et al* Gonadotropin releasing hormone Molecular and c biology, physiology, and clinical applications *Fed Proc*, *43* 22 1984

Cowie, A T, *et al* Hormonal Control of Lactation. New York, Springe Verlag, 1980

Cummings, A M, and Yochim, J M Differentiation of the uterus preparation for gestation A model for the action of progesterone *Theor Biol*, *106* 353, 1984

Fisher, D A Fetal endocrinology Endocrine disease and pregnancy DeGroot, L J, et al (eds) Endocrinology Vol 3 New York, Gru & Stratton, 1979, p 1649

Frantz, A G Prolactin *In* DeGroot, L J, *et al* (eds.) Endocrinolo Vol 1 New York, Grune & Stratton, 1979, p 153

Grant, N F, and Worley, R Hypertension in Pregnancy Concepts a Management New York, Appleton Century-Crofts, 1980

Habenicht, U F Hormonal Regulation of Testicular Descent N York Springer-Verlag, 1983

Hankins, G D, *et al* Myocardial infarction during pregnancy A revie *Obstet Gynecol*, *65* 139, 1985

Jaffe, R B The endocrinology of pregnancy *In* Yen S S C, and Jaf R B (eds) Reproductive Endocrinology Philadelphia, W B Saund Co, 1978, p 521

Leong D A *et al* Neuroendocrine control of prolactin secretion. *Ann Rev Physiol* *45* 109 1983

Li, C H (ed) The Chemistry of Prolactin New York, Academic Pre 1980

Lindheimer M D, and Katz A I Pathophysiology of preeclamps *Annu Rev Med*, *32* 273 1981

Meschia, G Circulation to female reproductive organs *In* Shepherd, T and Abboud F M (eds) Handbook of Physiology Sec 2, Vol I Bethesda American Physiological Society, 1983, p. 241

Munro H N *et al* The placenta in nutrition *Annu Rev Nutr*, *2* 1983

Nathanielsz P W Endocrine mechanisms of parturition *Annu R Physiol* *40* 411 1978

Ochiai, K (ed) Endocrine Correlates of Reproduction New Yo Springer-Verlag 1984

Pitkin, R M Calcium metabolism in pregnancy and the perina period A review *Am J Obstet Gynecol*, *151* 99, 1985

Sampson, D A and Jansen, G R Protein and energy nutrition duru lactation *Annu Rev Nutr*, *4* 43, 1984

Sherman, A I The role of prostaglandins in obstetrics gynecology *Sp Top Endocrinol Metab*, *6* 141 1984

Shiu R P C, and Friesen, H G Mechanisms of action of prolactin the control of mammary gland function *Annu Rev Physiol*, *42* 8 1980

Simpson, E R, and MacDonald, P C Endocrine physiology of t placenta *Annu Rev Physiol* *43* 163, 1981

Smith, M S Role of prolactin in mammalian reproduction In Gree R O (ed) International Review of Physiology Reproductive Phy ology III Vol 22 Baltimore, University Park Press, p 249

Steinberger, E Genetics, anatomy, fetal endocrinology *In* DeGroot, J, *et al* (eds) Endocrinology Vol 3 New York Grune & Stratto 1979, p 1309

Sutherland, H W, and Stowers, J M (eds) Carbohydrate Metabolis in Pregnancy and the Newborn 1979 New York, Springer-Verla 1979

Szeto, H H Pharmacokinetics in the ovine maternal fetal unit *Ann Rev Pharmacol Toxicol* *22* 221, 1982

Thorburn G D, and Challis, J R G Endocrine control of parturitio *Physiol Rev*, *59* 863, 1979

Tienhoven, A V Reproductive Physiology of Vertebrates Ithaca, Co nell University Press, 1982

Vorherr, H (eds) Human Lactation, New York, Grune & Stratto 1979

Wilson, J D Sexual differentiation *Annu Rev Physiol*, *40* 279, 197

Yen, S S C, and Jaffe, R B (eds) Reproductive Endocrinolog Physiology, Pathophysiology, and Clinical Management Philadelph W B Saunders Co, 1978

83

Special Features of Fetal and Neonatal Physiology

A complete discussion of fetal development, functioning of the child immediately after birth, and growth and development through the early years of life lies within the province of formal courses in obstetrics and pediatrics. However, many aspects of these are strictly physiological problems, some of which relate to the physiological principles that we have discussed for the adult and some of which are peculiar to the infant itself. The present chapter delineates and discusses some of the most important of these special problems

GROWTH AND FUNCTIONAL DEVELOPMENT OF THE FETUS

Early development of the placenta and of the fetal membranes occurs far more rapidly than development of the fetus itself. During the first two to three weeks the fetus remains almost microscopic in size, but thereafter, as illustrated in Figure 83–1, the dimensions of the fetus increase almost in proportion to age. At 12 weeks the length of the fetus is approximately 10 cm; at 20 weeks, 25 cm; and at term (40 weeks), 53 cm (about 21 inches). Because the weight of the fetus is proportional to the cube of the length, the weight increases *approximately in proportion to the cube of the* age of the fetus. Note in Figure 83–1 that the weight of the fetus remains almost nothing during the first months and reaches 1 pound only at five and a half months of gestation Then, during the last trimester of pregnancy, the fetus gains tremendously so that two months prior to birth the weight averages 3 pounds, one month prior to birth 4 5 pounds, and at birth 7 pounds, this birth weight varying from as low as 4.5 pounds to as high as 11 pounds in completely normal infants with completely normal gestational periods

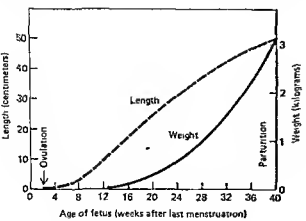


Figure 83–1. Growth of the fetus.

DEVELOPMENT OF THE ORGAN SYSTEMS

Within one month after fertilization of the ovum all the different organs of the fetus have already been "blocked out," and during the next two to three months, most of the details of the different organs are established. Beyond the fourth month, the organs of the fetus are grossly the same as those of the newborn child. However, cellular development of these structures is *usually far from complete at this time and requires the* full remaining five months of pregnancy for complete development. Even at birth certain structures, particularly the nervous system, the kidneys, and the liver, still lack full development, which will be discussed in more detail later in the chapter

The Circulatory System

The human heart begins beating during the fourth week following fertilization, contracting at the rate of about 65 beats per minute This *increases steadily as* the fetus grows and reaches a rate of approximately 140 beats per minute immediately before birth.

Formation of Blood Cells. Nucleated red blood cells begin to be formed in the yolk sac and mesothelial layers *of the placenta at about the third week of fetal* development. This is followed a week later by the formation of non-nucleated red blood cells by the fetal mesenchyme and by the endothelium of the fetal blood vessels. Then at approximately six weeks, the liver begins to form blood cells, and in the third month the spleen and other lymphoid tissues of the body also begin forming blood cells Finally, from approximately the third month on, the bone marrow also forms red and white blood cells. During the midportion of fetal life, the extramarrow areas are the major sources of the fetus' blood cells,

during the latter three months of fetal life, the bone marrow gradually takes over while the other blood cell–forming areas lose completely their ability to form blood cells.

The Respiratory System

Obviously, respiration cannot occur during fetal life. However, respiratory movements do take place beginning at the end of the first trimester of pregnancy. Tactile stimuli or fetal asphyxia especially cause respiratory movements

However, during the latter three to four months of pregnanry, the respiratory movements of the fetus are mainly inhibited, for reasons unknown This could possibly result from (1) special chemical conditions in the body fluids of the fetus, (2) the presence of fluid in the fetal lungs, or (3) other possible unexplored stimuli.

The inhibition of respiration during the latter months of fetal life prevents filling of the lungs with debris from the meconium excreted by the gastrointestinal tract into the amniotic fluid Also, fluid is secreted into the lungs by the alveolar epithelium up until the moment of birth, thus filling the pulmonary spaces with this clean secretion

The Nervous System

Most of the skin reflexes of the fetus are present by the third to fourth months of pregnancy However, most higher functions of the central nervous system that involve the cerebral cortex are still undeveloped even at birth Indeed, myelinization of some major tracts of the central nervous system becomes complete only after approximately a year of postnatal life

The Gastrointestinal Tract

By midpregnanry the fetus ingests and absorbs large quantities of amniotic fluid, and, during the last two to three months, gastrointestinal function approaches that of the normal newborn infant Small quantities of *meconium* are continually formed in the gastrointestinal tract and excreted from the bowels into the amniotic fluid Meconium is composed partly of unabsorbed residue of amniotic fluid and partly of excretory products from the gastrointestinal mucosa and glands.

The Kidneys

The fetal kidneys are capable of excreting urine during at least the latter half of pregnancy, and urination occurs normally *in utero* However, the renal control systems for regulation of extracellular fluid electrolyte balances and especially acid-base balance are almost nonexistent until after midfetal life and do not reach full development until a few months after birth

Fetal Metabolism

The fetus utilizes mainly glucose for energy, and it has a high rate of storage of fat and protein, much if not most of the fat being synthesized from glucose rather than being absorbed from the mother's blood Aside from these generalities there are some special problems of

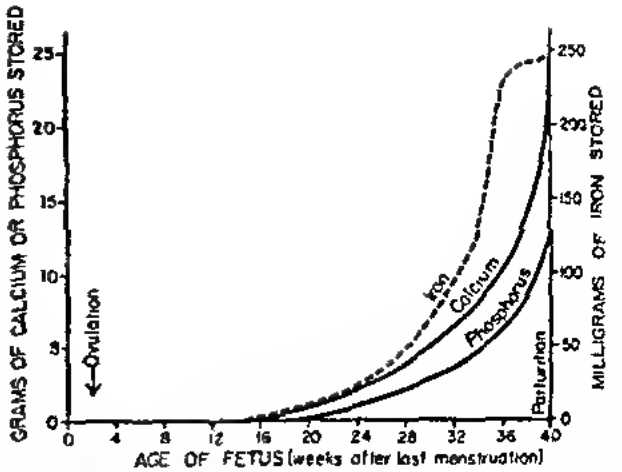


Figure 83–2. Calcium, phosphorus, and iron storage in the fetus at different stages of gestation.

fetal metabolism in relation to calcium, phosphate, iron, and some vitamins, discussed next.

Metabolism of Calcium and Phosphate. Figure 83–2 illustrates the rates of calcium and phosphate accumulation in the fetus, showing that approximately 22.5 grams of calcium and 13.5 grams of phosphorus are accumulated in the average fetus during gestation About half of this accumulates during the last four weeks of gestation, which is also coincident with the period of rapid ossification of the fetal bones as well as with the period of rapid weight gain of the fetus.

During the earlier part of fetal life, the bones are relatively unossified and have mainly a cartilaginous matrix Indeed, x-ray pictures ordinarily will not show any ossification until approximately the fourth month of pregnancy

Note especially that the total amounts of calcium and phosphate needed by the fetus during gestation represent only about one fifteenth the quantities of these substances in the mother's bones. Therefore, this is a minimal drain from the mother. However, a much greater drain occurs after birth during lactation.

Accumulation of Iron. Figure 83–2 shows that iron accumulates in the fetus somewhat more rapidly than calcium and phosphates. Most of the iron is in the form of hemoglobin, which begins to be formed as early as the third week following fertilization of the ovum.

Small amounts of iron are concentrated in the progestational endometrium even prior to implantation of the ovum; this iron is ingested into the embryo by the trophoblastic cells for early formation of the red blood cells

Approximately one third of the iron in a fully developed fetus is normally stored in the liver. This iron can then be used for several months after birth by the newborn infant for formation of additional hemoglobin.

Utilization and Storage of Vitamins. The fetus needs vitamins equally as much as the adult and in some instances to a far greater extent. In general, the vitamins function the same in the fetus as in the adult as discussed in Chapter 73. Special functions of several vitamins should be mentioned, however.

The B vitamins, especially vitamin B_{12} and folic acid, are necessary for formation of red blood cells and also for overall growth of the fetus.

Vitamin C is necessary for appropriate formation of

intercellular substances, especially the bone matrix and fibers of connective tissue.

Vitamin D probably is needed for normal bone growth in the fetus, but even more importantly the mother needs it for adequate absorption of calcium from her gastrointestinal tract. If the mother has plenty of this vitamin in her body fluids, large quantities will be stored by the fetal liver to be used by the newborn child for several months after birth.

Vitamin E, though its precise function is unknown, is necessary for normal development of the early embryo. In its absence in experimental animals, spontaneous abortion usually occurs at an early age.

Vitamin K is used by the fetal liver for formation of factor VII, prothrombin, and several other blood coagulation factors. When vitamin K is insufficient in the mother, factor VII and prothrombin become deficient in the child as well as in the mother. Since most vitamin K absorbed into the body is formed by bacterial action in the colon, the newborn child has no adequate source of vitamin K for the first week or so of life—that is, until a normal colonic bacterial flora becomes established. Therefore, prenatal storage of at least small amounts of vitamin K is helpful in preventing hemorrhage—particularly in the brain when the head is traumatized by squeezing through the birth canal.

ADJUSTMENTS OF THE INFANT TO EXTRAUTERINE LIFE

ONSET OF BREATHING

The most obvious effect of birth on the baby is loss of the placental connection with the mother, and therefore loss of this means for metabolic support. And by far the most important immediate adjustment required of the infant is to begin breathing.

Cause of Breathing at Birth. Following completely normal delivery from a mother who has not been depressed by anesthetics, the child ordinarily begins to breathe immediately and has a completely normal respiratory rhythm from the outset. The promptness with which the fetus begins to breathe indicates that breathing is initiated by sudden exposure to the exterior world, probably resulting from a slightly asphyxiated state incident to the birth process but also from sensory impulses originating in the suddenly cooled skin. In an infant who does not breathe immediately, the body becomes progressively more hypoxic and hypercapnic, which provides additional stimulus to the respiratory center and usually causes breathing within a few seconds to a few minutes after birth.

Delayed and Abnormal Breathing at Birth—Danger of Hypoxia. If the mother has been depressed by a general anesthetic during delivery, which at least partially also anesthetizes the child, respiration is likely to be delayed for several minutes, thus illustrating the importance of using as little obstetrical anesthesia as feasible. Also, many infants who have had head trauma during delivery or who undergo very prolonged delivery are slow to breathe or sometimes will not breathe at all. This can result from two possible effects: First, in a few infants, intracranial hemorrhage or brain contusion causes a concussion syndrome with a greatly depressed respiratory center. Second, and probably much more important, prolonged fetal hypoxia during delivery also causes serious depression of the respiratory center. Hypoxia frequently occurs during delivery because of (a) compression of the umbilical cord; (b) premature separation of the placenta; (c) excessive contraction of the uterus, which cuts off the blood flow to the placenta; or (d) excessive anesthesia of the mother, which depresses the oxygenation even of her blood.

Degree of Hypoxia That an Infant Can Tolerate. In the adult, failure to breathe for only four minutes often causes death, but a newborn infant often survives as long as 10 to 15 minutes of failure to breathe after birth. Unfortunately, though, permanent and very evident brain impairment often ensues if breathing is delayed more than 8 to 10 minutes. Indeed, actual lesions develop mainly in the thalamus, the inferior colliculi, and in other brain stem areas, thus affecting many of the motor functions of the body.

Expansion of the Lungs at Birth. At birth, the walls of the alveoli are kept collapsed by the surface tension of the viscid fluid that fills them. More than 25 mm Hg of negative pressure is required to oppose the effects of this surface tension and therefore to open the alveoli for the first time. But once the alveoli are open, further respiration can be effected with relatively weak respiratory movements. Fortunately, the first inspirations of the newborn infant are extremely powerful, usually capable of creating as much as 60 mm Hg negative pressure in the intrapleural space

Figure 83–3 illustrates the tremendous forces required to open the lungs at the onset of breathing. To the left is shown the pressure-volume curve (compliance curve) for the first breath after birth. Observe, first, the lowermost curve, which shows that the lungs hardly expand at all until the negative pressure has reached −40 cm water (−30 mm Hg). Then, as the negative pressure increases to −60 cm water, about 40 ml of air enters the lungs. To deflate the lungs, considerable positive pressure is required because of the viscous resistance offered by the fluid in the bronchioles.

Note that the second breath is much easier However, breathing does not become completely normal until about 40 minutes after birth, as shown by the third compliance curve, the shape of which compares favorably with that for the normal adult, as shown in Chapter 39.

Respiratory Distress Syndrome. A small number of infants, especially premature infants and infants born of diabetic mothers, develop severe respiratory distress a few hours to several days following birth and frequently succumb within the next day or so. The alveoli of these infants at death contain large quantities of proteinaceous fluid, almost as if pure plasma had leaked out of the capillaries into the alveoli. The fluid also contains desquamated alveolar epithelial cells. This con-

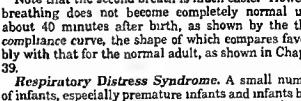


Figure 83–3. Pressure-volume curves of the lungs (compliance curves) of a newborn baby immediately after birth, showing *(a)* the extreme forces required for breathing during the first two breaths of life and *(b)* development of a nearly normal compliance within 40 minutes after birth. (From Smith. *Sci. Amer.*, 209:32, 1963, © 1963 by Scientific American, Inc. All rights reserved.)

dition is also called *hyaline membrane disease* because microscopic slides of the lung show this alveolar material to look like a hyaline membrane.

Unfortunately, the cause of the respiratory distress syndrome is not certain. However, one of the most characteristic findings is failure to secrete adequate quantities of *surfactant*, a substance normally secreted into the alveoli that decreases the surface tension of the alveolar fluid, therefore allowing the alveoli to open easily. Though this failure to secrete surfactant may be secondary to other pathological changes in the pulmonary membrane, some research workers believe it to be the basic cause of the disease and that the pathology is the secondary result. The surfactant-secreting cells (the type II alveolar epithelial cells) do not begin to secrete surfactant until the last one to three months of gestation. Therefore, many premature babies and some full-term babies are born without the capability of secreting surfactant, which therefore causes both a collapse tendency of the lungs and development of pulmonary edema. The role of surfactant in preventing these effects was discussed in Chapter 39.

CIRCULATORY READJUSTMENTS AT BIRTH

Almost as important as the onset of breathing at birth are the immediate circulatory adjustments that allow adequate blood flow through the lungs. Also, circulatory adjustments during the first few hours of life shunt more and more blood through the liver as well. To describe these readjustments we must first consider briefly the anatomical structure of the fetal circulation.

Specific Anatomic Structure of the Fetal Circulation. Because the lungs are mainly nonfunctional during fetal life and because the liver is only partially functional, it is not necessary for the fetal heart to pump much blood through either the lungs or the liver. On the other hand, the fetal heart must pump large quantities of blood through the placenta. Therefore, special anatomical arrangements cause the fetal circulatory system to operate considerably differently from that of the adult. First, as illustrated in Figure 83–4, blood returning from the placenta through the umbilical vein passes through the *ductus venosus*, mainly bypassing the liver. Then, most of the blood entering the right atrium from the inferior vena cava is directed in a straight pathway across the posterior aspect of the right atrium and thence through the *foramen ovale* directly into the left atrium. Thus, this well-oxygenated blood from the placenta enters the left side of the heart rather than the right side and is pumped by the left ventricle mainly into the vessels of the head and forelimbs.

The blood entering the right atrium from the superior vena cava is directed downward through the tricuspid valve into the right ventricle. This blood is mainly deoxygenated blood from the head region of the fetus, and it is pumped by the right ventricle into the pulmonary artery, then mainly through the *ductus arteriosus* into the descending aorta and through the two umbilical arteries into the placenta where the deoxygenated blood becomes oxygenated.

Figure 83–5 gives the relative percentages of the total blood pumped by the heart that passes through the different vascular circuits of the fetus. This figure shows that 55 per cent of all the blood goes through the placenta, leaving only 45 per cent to pass through all of the fetus. Furthermore, during fetal life, only 12 per cent of the blood flows through the lungs; immediately after birth three to four times this much must flow through the lungs, indicating a severalfold increase at birth.

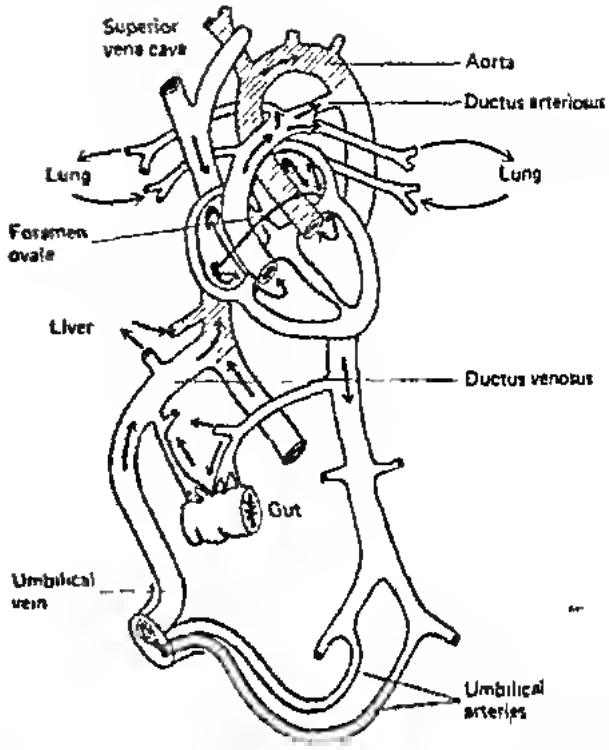


Figure 83–4. Organization of the fetal circulation. (Modified from Arey: Developmental Anatomy, 7th ed. Philadelphia, W. B. Saunders Company, 1974.)

Changes in the Fetal Circulation at Birth. The basic changes in the fetal circulation at birth were discussed in Chapter 27 in relation to congenital anomalies of the ductus arteriosus and foramen ovale that persist throughout life. Briefly, these changes are the following:

Primary Changes in Pulmonary and Systemic Vascular Resistance at Birth. The primary changes

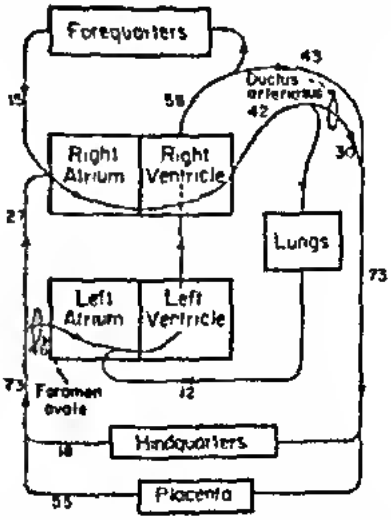


Figure 83–5. Diagram of the fetal circulatory system, showing relative distribution of blood flow to the different vascular areas. The numerals represent per cent of the total cardiac output flowing through the particular area.

in the circulation at birth are first, loss of the tremendous blood flow through the placenta, which *approximately doubles the systemic vascular resistance at birth.* This obviously *increases the aortic pressure* as well as the pressures in the left ventricle and left atrium.

Second, the *pulmonary vascular resistance greatly decreases* as a result of expansion of the lungs. In the unexpanded fetal lungs, the blood vessels are compressed because of the small volume of the lungs. Immediately upon expansion these vessels are no longer compressed, and the resistance to blood flow decreases several-fold. Also, in fetal life the hypoxia of the lungs causes considerable tonic vasoconstriction of the lung blood vessels, but vasodilation takes place when aeration of the lungs eliminates the hypoxia. All these changes together reduce the resistance to blood flow through the lungs as much as fivefold, which obviously *reduces the pulmonary arterial pressure*, the right ventricular pressure, and the right atrial pressure.

Closure of the Foramen Ovale. The *low right atrial pressure* and the *high left atrial pressure* that occur secondarily to the changes in pulmonary and systemic resistances at birth cause a tendency for blood to flow backward from the left atrium into the right atrium rather than in the other direction, as occurred during fetal life. Consequently, the small valve that lies over the foramen ovale on the left side of the atrial septum closes over this opening, thereby preventing further flow through the foramen ovale. In two-thirds of all persons the valve becomes adherent over the foramen ovale within a few months to a few years and forms a permanent closure. But, even if permanent closure does not occur, the left atrial pressure throughout life remains 2 to 4 mm Hg greater than the right atrial pressure, and the back pressure keeps the valve closed.

Closure of the Ductus Arteriosus. The ductus arteriosus also closes, but for different reasons. First, the increased systemic resistance *elevates the aortic pressure* while the decreased pulmonary resistance *reduces the pulmonary arterial pressure.* As a consequence, within a few hours after birth, blood begins to flow backward from the aorta into the pulmonary artery rather than in the other direction as in fetal life However, after only a few hours the muscular wall of the ductus arteriosus constricts markedly, and within one to eight days the constriction is sufficient to stop all blood flow. This is called *functional closure* of the ductus arteriosus. Then, during the next one to four months the ductus arteriosus ordinarily becomes anatomically *occluded* by growth of fibrous tissue into its lumen.

The cause of ductus closure almost certainly relates to the increased oxygenation of the blood flowing through the ductus. In fetal life the Po_2 of the ductus blood is *only 15 to 20 mm Hg*, but *it increases to about* 100 mm Hg within a few hours after birth. Furthermore, many experiments have shown that the degree of contraction of the smooth muscle in the ductus wall is highly related to the availability of oxygen.

In one of several thousand infants the ductus fails to close, resulting in a *patent ductus arteriosus*, the consequences of which were discussed in Chapter 27.

Closure of the Ductus Venosus. In fetal life, the portal blood from the fetus' abdomen joins the blood from the umbilical vein and these together pass through the ductus venosus directly into the vena cava, thus bypassing the liver. During the first few hours after birth, blood flow through the umbilical vein ceases, but most of the portal blood still flows through the ductus venosus, with only a small amount passing through the channels of the liver. However, within one to three hours the muscular wall of the ductus venosus contracts strongly and closes this avenue of flow. As a consequence, the portal venous pressure rises from near 0 mm Hg to 6 to 10 mm Hg, which is enough to force blood flow through the liver sinuses. Although the ductus venosus almost never fails to close, unfortunately we know almost nothing about what causes the closure.

NUTRITION OF THE NEWBORN INFANT

The fetus obtains almost all its energy from glucose obtained from the mother's blood. Immediately after birth, the amount of glucose stored in the infant's body in the form of glycogen is sufficient to supply the infant's needs for only a few hours, and unfortunately the liver of the newborn infant is still far from functionally adequate at birth, which prevents significant gluconeogenesis Therefore, the infant's blood glucose concentration frequently falls the first day to as low as 30 to 40 mg/100 ml of plasma, less than half the value Fortunately, appropriate mechanisms are available for the infant to utilize stored fats and proteins for metabolism until mother's milk can be provided two to three days later

Special problems are also frequently associated with getting an adequate fluid supply to the newborn infant, because the infant's rate of body fluid turnover averages seven times that of an adult, and the mother's milk supply requires several days to develop Ordinarily, the infant's weight decreases 5 to 10 per cent and sometimes as much as 20 per cent within the first two to three days of life Most of this weight loss is loss of fluid rather than of body solids

SPECIAL FUNCTIONAL PROBLEMS IN THE NEONATAL INFANT

The most important characteristic of the newborn infant is instability of the various hormonal and neurogenic control systems. This results partly from the immature development of the different organs of the body and partly from the fact that the control systems simply have not become adjusted to the completely new way of life.

The Respiratory System. The normal rate of respiration in the newborn is about 40 breaths per minute, and tidal air with each breath averages 16 ml. This *gives a total minute respiratory volume of 640 ml per* minute, which is about two times as great in relation to the body weight as that of an adult. *The functional residual capacity of the infant is only half that of an adult in relation to body weight.* This allows cyclic changes in blood gas concentration when the respiration becomes slowed because it is the residual air in the lungs that smooths out the blood gas variations in the adult.

Blood Volume. The blood volume of a newborn infant immediately after birth averages about 300 ml, but, if the infant is left attached to the placenta for a few minutes after birth or if the umbilical cord is stripped to force blood out of its vessels into the baby, an addi-

tional 75 ml of blood enters the infant to make a total of 375 ml. Then, during the ensuing few hours, fluid is lost into the tissue spaces from this blood, which increases the hematocrit but returns the blood volume onre again to the normal value of about 300 ml. Some pediatricians feel that this extra blood volume in some instances causes mild pulmonary edema with some degree of respiratory distress.

Cardiac Output. The cardiac output of the newborn infant avemges 550 ml/minute, which, like respiration and body metabolism, is about two times as much in relation to body weight as in the adult. Occasionally a child is born with an especially low cardiac output caused by bemorrhage through the placental membrane into the mother's body prior to birth.

Arterial Pressure. The arterial pressure during the first day after birth averages about 70/50; this increases slowly during the next several months to approximately 90/60. Then there is a much slower rise during the subsequent years until the adult pressure of 120/80 is attained at adolescence.

Blood Characteristics. The red blood cell count in the newborn infant averages about 4 million per cubic millimeter If blood is stripped from the cord into the infant, tho red blood rell count rises en additional 0.5 to 0.75 million during the first few hours of life, giving a red blood cell count of about 4 75 million per eubic millimeter, as illustrated in Figure 83–6 Subsequent to this, however, few new red blood cells are formed in the infant during the first few weeks of life, presumably because tbe hypoxic stimulus of fetal life is no longer present to stimulate red cell production. Thus, as shown in Figure 83–6, the average red blood cell count falls to 3 25 million per cubic millimeter by about 8 to 10 weeks of age From that time on, increasing activity by the fetus provides the appropriato stimulus for returning the red blood cell count to normal within another two to three months

Immediately after birth, the white blood cell count of the infant is about 45,000 per cubic millimeter, which is about five times as great as that of the normal adult.

Neonatal Jaundice and Erythroblastosis Fetalis. Bilirubin formed in the fetus can cross the placenta and be excreted through the liver of the mother, but immediately after birth the only means for ridding the infant of bilirubin is through the infant's own liver, which, for the first week or so of life, still functions poorly and is

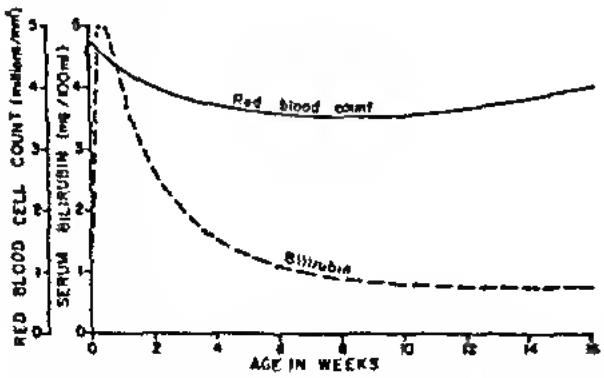


Figure 83–6. Changes in the red blood cell count and in the serum bilirubin concentration during the first 16 weeks of life, showing "physiologic anemia" at 6 to 12 weeks of life and "physiologic hyperbilirubinemia" during the first 2 weeks of life

incapable of conjugating significant quantities of bilirubin with glucuronic acid for excretion into the bile. Consequently, the plasma bilirubin concentration rises from a normal value of less than 1 mg/100 ml to an average of 5 mg/100 ml during the first three days of life, then gradually falls back to normal as the liver becomes functional. This condition, called *physiological hyperbilirubinemia*, is illustrated in Figure 83–6, and it is associated with a mild *jaundice* (yellowness) of the infant's skin and especially of the sclerae of its eyes

However, by far the most important cause of serious neonatal jaundice is *erythroblastosis fetalis*, which was discussed in detail in Chapter 4 in relation to Rh factor incompatibility between the infant and mother. Briefly, the baby usually inherits an Rh positive trait from the father while the mother is Rh negative. The mother then becomes immunized against the fetus' blood, and her antibodies in turn destroy the infant's red blood cells, releasing extreme quantities of bilirubin into the plasma Before the advent of modern obstetrical methods, this condition occurred either mildly or seriously in 1 of every 50 to 100 newborn infants.

Fluid Balance, Acid-Base Balance, and Renal Function. The rate of fluid intake and fluid excretion in the infant is seven times as great in relation to weight as in the adult, which means that even a slight alteration of fluid balance can cause rapidly developing abnormalities. Second, the rate of metabolism in the infant is two times as great in relation to body mass as in the adult, which means that two times as much acid is normally formed, which leads to a tendency toward acidosis in the infant. Third, functional development of the kidneys is not complete until the end of approximately the first month of life. For instanre, the kidneys of the newborn can concentrate urine to only 1.5 times the osmolality of the plesme instead of the narmal three to four times in the adult.

Therefore, considering the immaturity of the kidney, together with the marked fluid turnover in the infant and rapid formation of acid, one can readily understand that among the most important problems of infancy are acidosis, dehydration, and some instances of overhydration.

Liver Function. During the first few days of life, liver function may be quite deficient, as evidenced by the following effects:

1. The liver of the newborn conjugates bilirubin with glucuronic acid poorly and therefore excretes bilirubin only slightly during the first few days of life.

2 The liver of the newborn is deficient in forming plasma proteins, so that the plasma protein concentration falls to 1 gram per cent less than that for older children. Occasionally, the protein concentration falls so low that the infant actually develops hypoproteinemic edema.

3. The gluconeogenetic function of the liver is particularly deficient. As a result, the blood glucose level of the unfed newborn infant falls to about 30 to 40 mg per cent, and the infant must depend on its stored fats for energy until feeding can occur.

4. The liver of the newborn usually also forms too little of the factors needed for normal blood coagulation.

Digestion, Absorption, and Metabolism of Energy Foods. In general, the ability of the newborn infant to digest, absorb, and metabolize foods is not different from that of the older child, with the following three exceptions:

First, secretion of pancreatic amylase in the newborn infant is deficient so that the infant utilizes starches less adequately than do older children.

Second, absorption of fats from the gastrointestinal tract is somewhat less then that in the older child. Consequently, milk with a high fat content, such as cow's milk, is frequently inadequately utilized.

Third, because the liver functions imperfectly during at least the first week of life, the glucose concentration in the blood is unstable and also low.

The neonate is especially capable of synthesizing proteins and thereby storing nitrogen. Indeed, with a completely adequate diet, as much as 90 per cent of the ingested amino acids are utilized for formation of body proteins. This is a much higher percentage than in adults.

Metabolic Rate and Body Temperature. The normal metabolic rate of the neonate in relation to body weight is about two times that of the adult, which accounts also for the two times as great cardiac output and two times as great minute respiratory volume in the infant.

However, since the body surface area is very large in relation to body mass, heat is readily lost from the body. As a result, the body temperature of the newborn infant, particularly of premature infants, falls easily. Figure 83–7 shows that the body temperature of even the normal infant falls several degrees during the first few hours after birth, but returns to normal in seven to ten hours. Still, the body temperature regulatory mechanisms remain poor during the early days of life, allowing marked deviations in temperature, which are also illustrated in Figure 83–7.

Nutritional Needs During the Early Weeks of Life. At birth, a newborn infant is usually in complete nutritional balance, provided the mother has had an adequate diet. Furthermore, function of the gastrointestinal system is usually more than adequate to digest and assimilate all the nutritional needs of the infant if these are provided in the diet. However, three specific problems do occur in the early nutrition of the infant.

Need for Calcium and Vitamin D. The newborn infant is in a stage of rapid ossification of its bones at birth so that a ready supply of calcium throughout infancy is needed. This is ordinarily supplied adequately by the usual diet of milk. Yet absorption of calcium by the gastrointestinal tract is poor in the absence of vitamin D. Therefore, the vitamin D–deficient infant can develop severe rickets in only a few weeks. This is particularly true in premature babies since their gastrointestinal tracts absorb calcium even less effectively than those of normal infants.

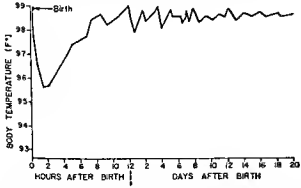


Figure 83–7. Fall in body temperature of the infant immediately after birth, and instability of body temperature during the first few days of life.

Necessity for Iron in the Diet. If the mother has had adequate amounts of iron in her diet, the liver of the infant usually has stored enough iron to keep forming blood cells for four to six months after birth. But if the mother has had insufficient iron in her diet, severe anemia is likely to occur in the infant after about three months of life. To prevent this possibility, early feeding of egg yolk, which contains reasonably large quantities of iron, or administration of iron in some other form is desirable by the second or third month of life.

Vitamin C Deficiency in Infants. Ascorbic acid (vitamin C) is not stored in significant quantities in the fetal tissues; yet it is required for proper formation of cartilage, bone, and other intercellular structures of the infant. Furthermore, milk has poor supplies of ascorbic acid, especially cow's milk, which has only one fourth as much as mother's milk. For this reason, orange juice or other sources of ascorbic acid are usually prescribed by the third week of life.

Immunity. Fortunately, the neonate inherits much immunity from the mother because many antibodies diffuse from the mother's blood through the placenta into the fetus. However, the neonate does not form antibodies of its own to a significant extent. By the end of the first month, the baby's gamma globulins, which contain the antibodies, have decreased to less than one half the original level, with corresponding decrease in immunity. Thereafter, the baby's own immunization processes begin to form antibodies, and the gamma globulin concentration returns essentially to normal by the age of 12 to 20 months.

Despite the decrease in gamma globulins soon after birth, the antibodies inherited from the mother still protect the infant for about six months against most major childhood infectious diseases, including diphtheria, measles, and polio. Therefore, immunization against these diseases before six months is usually unnecessary. On the other hand, the inherited antibodies against whooping cough are normally insufficient to protect the neonate; therefore, for full safety the infant requires immunization against this disease within the first month or so of life.

Allergy. Fortunately, the newborn infant is rarely subject to allergy. Several months later, however, when the neonate's own antibodies first begin to form, extreme allergic states can develop, often resulting in serious eczema, gastrointestinal abnormalities, or even anaphylaxis. As the child grows older and develops still higher degrees of immunity, these allergic manifestations usually disappear. This relationship of immunity to allergy was discussed in Chapter 6.

Endocrine Problems. Ordinarily the endocrine system of the infant is highly developed at birth, and the infant rarely exhibits any immediate endocrine abnormalities. However, there are special instances in which endocrinology of infancy is important.

1. If a pregnant mother bearing a female child is treated with an androgenic hormone or if she develops an androgenic tumor during pregnancy, the child will be born with a high degree of masculinization of her sexual organs, thus resulting in a type of *hermaphroditism.*

2. The sex hormones secreted by the placenta and by the mother's glands during pregnancy occasionally cause the neonate's breasts to form milk during the first few

safe with up to 40 per cent oxygen, but some child physiologists believe that complete safety can be achieved only at normal oxygen concentration.

GROWTH AND DEVELOPMENT OF THE CHILD

The major physiological problems of the child beyond the neonatal period are related to special metabolic needs for growth, which have been fully covered in the sections on metabolism and endocrinology

Figure 83–8 illustrates the changes in heights of boys and girls from the time of birth until the age of 20 years. Note especially that these parallel each other almost exactly until the end of the first decade of life Between the ages of 11 and 13 years the female estrogens cause rapid growth but early uniting of the epiphyses at about the 14th to 16th year of life, so that growth in height ceases. This contrasts with the effect of testosterone in the male, which causes growth at a slightly later age—mainly between ages 13 and 17 years. The male, however, undergoes much more prolonged growth so that his final height is considerably greater than that of the female.

BEHAVIORAL GROWTH

Behavioral growth is principally a problem of maturity of the nervous system. Here, it is extremely difficult to dissociate maturity of the anatomical structures of the nervous system from maturity caused by training Anatomical studies show that certain major tracts in the central nervous system are not completely myelinated until the end of the first year of life For this reason it is frequently stated that the nervous system is not fully functional at birth The brain cortex and its associated mechanisms such as vision seem to require several months after birth for significant functional development.

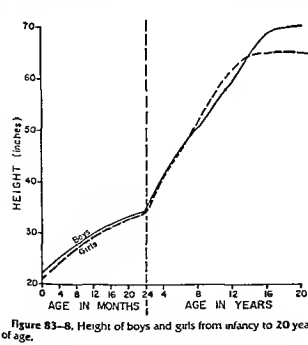


Figure 83–8. Height of boys and girls from infancy to 20 years of age.

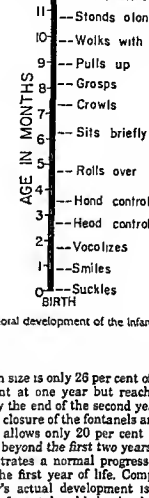


Figure 83–9. Behavioral development of the infant during the *first year of life*

At birth, the brain size is only 26 per cent of the adult size and 55 per cent at one year but reaches almost adult proportions by the end of the second year. This is also associated with closure of the fontanels and sutures of the skull, which allows only 20 per cent additional *growth of the brain beyond the first two years of life.*

Figure 83–9 illustrates a normal progress chart for the infant during the first year of life. Comparison of this with the baby's actual development is used for clinical assessment of mental and behavioral growth.

REFERENCES

Arey, L. B Developmental Anatomy. *7th Ed., Revised Philadelphia,* W B Saunders Co, 1974

Aynsley-Green, A Metabolic and endocrine interrelations in the human fetus and neonate *Am J Clin Nutr, 41* 399, 1985

Babson, S G, *et al* Diagnosis and Management of the Fetus and Neonate at Risk A Guide for Team Care St. Louis, C V Mosby, 1979

Battaglia, F C, and Meschia, G Principal substrates of fetal metabolism *Physiol Rev, 58* 499, 1978

Buggers, J D, and Borland, R M. Physiological aspects of growth and development of the preimplantation mammalian embryo *Annu Rev Physiol, 38* 95, 1976

Bessonnette, J M, *et al* Regulation of cerebral blood flow in the fetus *J Dev Physiol, 6* 275, 1984

Bradley, R. M, and Mistretta, C M Fetal sensory receptors. *Physiol Rev, 55*.352, 1975

Caccari, E, *et al* Obesity in Childhood New York, Academic Press, 1978

Cockburn, F, and Drillien, C M Neonatal Medicine Philadelphia, J B. Lippincott Co, 1975

Dallman, P R Anemia of prematurity *Annu. Rev Med, 32* 143, 1981

Daughaday, W H, *et al.* The regulation of growth by endocrines. *Annu Rev Physiol, 37*.211, 1975

Davis, J A., and Dobbind, J Scientific Foundations of Pediatrics. Philadelphia. W B Saunders Co, 1975

Durnin, J V.. Energy balance in childhood and adolescence. *Proc Nutr. Soc, 43*.271, 1984

Falkner, F, and Tanner, J M (eds) Human Growth. New York, Plenum Press, 1978.

Fanaroff, A. A Behrman's Neonatal Perinatal Medicine Diseases of the Fetus and Infant. St. Louis, C V. Mosby, 1983

Fenichel, G M: Neonatal Neurology New York, Churchill Livingstone, 1980
Fomon, S J Infant Nutrition, 2nd Ed Philadelphia, W. B Saunders Co, 1975
Gardner L. I Endocrine and Genetic Diseases of Childhood and Adolescence 2nd Ed Philadelphia, W B Saunders Co., 1975
Gaull, G E *et al* Significance of growth modulators in human milk *Pediatrics 75* 142, 1985
Gluck, L (ed) Intrauterine Asphyxia and the Developing Fetal Brain Chicago, Year Book Medical Publishers 1977
Godman M J, and Marquis, R M: Paedatric Cardiology Heart Disease in the Neonate New York, Churchill Livingstone, 1979
Grundmann, E, and Kirsten, W H (eds) Perinatal Pathology New York, Springer-Verlag, 1979
Hafez, E S E (ed) The Mammalian Fetus, Springfield, Ill, Charles C Thomas, 1975
Haller, J O, and Schneider, M Pediatric Ultrasound Chicago Year Book Medical Publishers, 1980
Harding, R Function of the larynx in the fetus and newborn *Annu Rev Physiol*, *46* 645, 1984
Haymond, M W, and Pagliara, A. S Endocrine and metabolic aspects of fuel homeostasis in the fetus and neonate *In* DeGroot, L. J, *et al* (eds) Endocrinology Vol 3 New York, Grune & Stratton, 1979, p 1779
Heymann M. A., *et al* Factors affecting changes in the neonatal systemic circulation *Annu Rev Physiol*, *43* 371, 1981
Jellife D B, and Jellife, E F P (eds) Nutrition and Growth, New York, Plenum Press, 1979
King D B, and May, J D Thyroidal influence on body growth *J Exp Zool*, *233* 453 1984
Klaus, M H, and Fanaroff, A A Care of the High Risk Neonate Philadelphia, W B Saunders Co, 1979
Kumar, S, and Rathi, M (eds) Perinatal Medicine Clinical and Biochemical Aspects of the Evaluation, Diagnosis and Management of the Fetus and Newborn New York, Pergamon Press, 1978
Lauer, R. M, and Shekelle, R B (eds) Childhood Prevention of Atherosclerosis and Hypertension New York, Raven Press, 1980
Lotgering, F K, *et al* Maternal and fetal responses to exercise during pregnancy *Physiol Rev*, *65* 1, 1985
Lough, M D, *et al* (eds.) Newborn Respiratory Care Chicago, Year Book Medical Publishers, 1979
MacLean Pediatric Nutrition in Clinical Practice Menlo Park, Addison, Wesley Publishing Co, 1984
Mender, H Introduction to the study of pre- and postnatal growth in humans A review *Am J Med. Genet*, *20* 63, 1985
Miller, D R, *et al* (eds) Smith's Blood Diseases of Infancy and Childhood St. Louis, C V Mosby, 1978
Miller, H C, and Merritt, T A Fetal Growth in Humans Chicago, Year Book Medical Publishers, 1979
Mott, J C and Walker, D W Neural and endocrine regulation of circulation in the fetus and newborn *In* Shepherd, J T, and Abboud, F. M (eds.) Handbook of Physiology, Sec 2, Vol III Bethesda American Physiological Society, 1983, p 837
Murray, R (ed) Respiratory Distress Syndrome. Edison, N. J., Medical Research Book Publishers, 1979
Nathan, D G, and Oski, F A (eds) Hematology of Infancy and Childhood. 2nd Ed Philadelphia, W. B Saunders Co, 1980
Nissley, S P, and Rechler, M M: Somatomedin/insulin-like growth factor tissue receptors *Clin Endocrinol. Metab*, *13* 43, 1984
Olley, P M, and Coceani, F. Prostaglandins and the ductus arteriosus *Annu Rev. Med*, *32* 375, 1981
Read, D J and Henderson-Smart, D J. Regulation of breathing in the newborn during different behavioral states *Annu. Rev Physiol* *46* 675, 1984
Rigatto, H: Control of ventilation in the newborn. *Annu Rev Physiol* *46* 661, 1984
Rosett, H L Alcohol and the Fetus A Clinical Perspective New York Oxford University Press, 1984
Rudolph A M Fetal and neonatal pulmonary circulation *Annu. Rev Physiol*, *41* 383, 1979
Scarpelli, E M, *et al* (eds.) Pulmonary Disease of the Fetus and Child Philadelphia, Lea & Febiger, 1978.
Schwartz, E (ed) Hemoglobinopathies in Children. Littleton, Mass PSG Publishing Co, 1979
Sinclair, D Human Growth after Birth. New York, Oxford University Press, 1978
Sinclair J C (ed) Temperature Regulation and Energy Metabolism the Newborn New York, Grune & Stratton, 1978
Smith, C A, and Nelson, N M The Physiology of the Newborn Infant Springfield Ill, Charles C Thomas, 1974
Stern L., *et al* (eds) Intensive Care in the Newborn, II New York Masson Publishing Co 1978
Strang, L. B Fetal and newborn lung *In* MTP International Review Science Physiology Vol 2 Baltimore, University Park Press, 197 p 31
Strang, L. B Neonatal Respiration Physiological and Clinical Studies Philadelphia, J B Lippincott Co, 1977
Sutherland, H W, and Stowers, J M (eds) Carbohydrate Metabolism in Pregnancy and the Newborn 1979 New York, Springer-Verlag 1979
Van Leeuwen, G Van Leeuwen's Newborn Medicine Chicago, Year Book Medical Publishers, 1979
Walker, D W Peripheral and central chemoreceptors in the fetus and newborn *Annu Rev Physiol*, *46* 687, 1984
Warshaw, J B (ed) Symposium on Fetal Disease Philadelphia, W B Saunders Co, 1979
Winick, M (ed) Nutrition, Pre- and Postnatal Development New York Plenum Press, 1979
Wolff, P H, and Ferber, P The development of behavior in human infants, premature and newborn *In* Cowan, W M, *et al* (eds.) Annu Rev Neurosci 2 291, 1979

XIV

SPORTS PHYSIOLOGY

84

Sports Physiology

It is fitting to end this text with a chapter on sports physiology because there are no normal stresses to which the body is exposed that even nearly approach the extreme stresses of heavy exercise In fact, if some of the extremes of exercise were continued for even slightly prolonged periods of time, they might easily be lethal Therefore, in the main, sports physiology is a discussion of the ultimate limits to which most of the bodily mechanisms can be stressed To give one simple example In a person who has extremely high fever, approaching the level of lethality, the body metabolism increases to about 100 per cent above normal. By comparison, the metabolism of the body during a marathon race increases to 2000 per cent above normal

The Female and the Male Athlete

Most of the quantitative data that will be given in this chapter will be for the young male adult, not because it is desirable to know only these values but because it is only in this class of athletes that relatively complete measurements have been made However, for those measurements that have been made in women, almost identically the same basic physiological principles apply equally as to men except for quantitative differences caused by differences in body size, body composition, and the presence or absence of the male sex hormone testosterone In general, most quantitative values for women—such as muscle strength, pulmonary ventilation, and cardiac output, all of which are related mainly to the muscle mass—will vary between two thirds and three quarters of the values recorded in men However, this does not translate into this much differential in athletic performance because the body size is correspondingly smaller A good indication of the relative performance capabilities of the female versus the male athlete comes from the relative times required for running the marathon race In a recent comparison, the top female performer had a running time about 12 per cent less than that of the top male performer. On the other hand, for some endurance events, women have proved to have capabilities superior to those of men As an example, the record for the two-way swim across the English channel is presently held by a woman, not a man Part of the reason for this has been reputed to be that the female has extra fat in her subcutaneous tissues to insulate her from the cold of the channel waters, but this certainly is not all of the reason, and it may be more the wishful thinking of the male ego rather than fact.

The hormonal differences between woman and man certainly account for a large part if not most of the differences in athletic performance *Testosterone* secreted by the male testicles has a powerful anabolic effect, which means that it causes greatly increased deposition of protein everywhere in the body, especially in the muscles In fact, even the man who participates in very little sports activity but who nevertheless is well-endowed with testosterone will have muscles that grow to sizes 40 per cent or more greater than those of his female counterpart and with a corresponding increase in strength. Thus, the man who begins to train for sports activity already has a running start on the female

The female sex hormone *estrogen* probably also accounts for some of the difference between female and male performance, though not nearly so much as the effect of testosterone Estrogen is known to increase the deposition of fat in the female, especially in certain tissues such as the breasts, the hips, and the subcutaneous tissue At least partly for this reason, the average nonathletic female has about 26 per cent body fat composition in contrast to the nonathletic male, who has about 15 per cent In marathon runners who have trained themselves to the least amount of excess fat, the male runner has about 4 per cent body fat composition and the female, 6 per cent. Thus, either in the untrained or the trained state, women usually average about 50 per cent more body fat than men This obviously is a detriment to the highest levels of athletic performance in those events in which performance depends upon speed or bodily strength, but on the other hand it could be an aid in grueling endurance athletic events that require the fat for energy

Estrogen plays another more insidious role in athletics, for it is the estrogen secreted by the female ovaries after puberty that makes the female stature smaller than that of the male Immediately after puberty, the surge in estrogen secretion causes a rapid spurt of growth that usually makes the postpubertal female grow more rapidly than her male counterpart. On the other hand, this growth is short-lived because the epiphyseal cartilages of the long bones, which is where the growth occurs, rapidly run their course and actually disappear, allowing the epiphyses to unite with the shafts of the long bones—therefore, no further growth in length. As a result, the female frequently reaches her full height at an age of perhaps 14 to 16 years, whereas the male may continue to grow until the age of 18 to 20 years This difference obviously looms large in most athletic events because the very design of athletic competition often gives the edge to those of greater body size

Finally, one cannot neglect the effect of the sex hor-

mones on temperament. There is no doubt that testosterone promotes aggressiveness and that estrogen is associated with a more mild temperament. Certainly a large part of competitive sports is the aggressive spirit that drives a person to maximum effort, often at the expense of judicious restraint.

THE MUSCLES IN EXERCISE

STRENGTH, POWER, AND ENDURANCE OF MUSCLES

The final common denominator in athletic events is what the muscles can do for you—what strength they can give when it is needed, what power they can achieve in the performance of work, and how long they can continue in their activity.

The strength of a muscle is determined mainly by its size, with a *maximum contractile force between 2.5 kg and 3.5 kg per cm²* of muscle cross-sectional area. Thus, the man who is well laced with testosterone and therefore has correspondingly enlarged muscles will be much stronger than those persons without the testosterone advantage. Also, the athlete who has hypertrophied his muscles through an exercise training program likewise will have increased muscle strength because of increased muscle size.

To give an example of muscle strength, a world-class weight lifter might have a quadriceps muscle with a cross-sectional area as great as 150 cm². This would translate into a maximum contractile strength of 525 kg (or 1155 lb), with all this force applied to the patellar tendon. Therefore, one can readily understand how it is possible for this tendon to be ruptured or actually to be avulsed from its insertion into the tibia below the knee. Also, when such *forces occur* in tendons that span a joint, similar forces are also applied to the surfaces of the joints, or sometimes to ligaments spanning the joints, thus accounting for such happenings as displaced cartilages, compression fractures about the joint, or torn ligaments.

Yet, to make matters still worsa, the *holding strength* of muscles is about 40 per cent greater than the contractile strength. That is, if a muscle is already contracted and a force then attempts to stretch out the muscle, this requires about 40 per cent more force than can be achieved by a shortening contraction. Therefore, the force of 525 kg calculated previously for the patellar tendon becomes 735 kg (1617 lb). This obviously further compounds the problems of the tendons, joints, and ligaments. It can also lead to internal tearing in the muscle itself. In fact, stretching out of a maximally contracted muscle is one of the best ways to insure the highest degree of muscle soreness.

The *power* of muscle contraction is different from muscle strength, for power is a measure of the amount of work that the muscle can perform in a given period of time. This is determined not only by the strength of muscle contraction but also by its *velocity of contraction* and the number of times that it contracts each minute.

Muscle power is generally measured in *kilogrm-meters (kg·m)/minute.* That is, a muscle that can lift a kilogram weight to a height of 1 meter or that can move some object laterally against a force of 1 kg for a distance of a meter in 1 minute is said to have a power of 1 kg-m/minute. The maximum power achievable by all the muscles in the body of a highly trained athlete with all the muscles working together is approximately the following:

First 10 to 15 seconds	7000 kg·m/minute
Next 1 minute	4000 kg·m/minute
Next half hour	1700 kg·m/minute

Thus, it is clear that a person has the capability of an extreme power surge for a short period of time, such as during a 100 meter dash that can be completed entirely within the first 10 seconds, whereas for long-term endurance events the power output of the muscles is only one fourth as great as during the initial power surge. Yet, this does not mean that one's athletic performance is four times as great during the initial power surge as it is for the next half hour, because the efficiency for translation of muscle power output into athletic performance is often much less during rapid activity than during less rapid but sustained activity. Thus, the velocity of the hundred meter dash is only 1.75 times as great as the velocity of the 30 minute race despite the fourfold difference in short-term versus long-term muscle power capability.

The final measure of muscle performance is *endurance.* This, to a great extent, depends on the nutritive support for the muscle—more than anything else on the amount of glycogen that has been stored in the muscle prior to the period of exercise. A person on a high carbohydrate diet stores far more glycogen in muscles than a person on either a mixed diet or a high fat diet. Therefore, endurance is greatly enhanced by a high carbohydrate diet. When athletes run at speeds typical for the marathon race, their endurance as measured by the time that they can sustain the race until complete exhaustion is approximately tha following:

High carbohydrate diet	240 minutes
Mixed diet	120 minutes
High fat diet	85 minutes

The corresponding amounts of glycogen stored in the muscle are approximately the following:

High carbohydrate diet	33 grams/kg of muscle
Mixed diet	17.5 grams/kg of muscle
High fat diet	6 grams/kg of muscle

THE MUSCLE METABOLIC SYSTEMS IN EXERCISE

The same basic metabolic systems are present in muscle as in all other parts of the body, these were discussed in detail in Chapters 67 through 73. However, special quantitative measures of the activities of three metobolic systems are exceedingly important in understanding the limits of physical activity. These are (1) the *phosphagen system,* (2) the *glycogen–lactic acid system,* and (3) the *aerobic system.*

The Phosphagen System

Adenosine Triphosphate. The basic source of energy for muscle contraction is adenosine triphosphate (ATP), which has the following basic formula:

$$\text{Adenosine}—PO_4\sim PO_4\sim PO_4$$

The bonds attaching the last two phosphate radicals to the molecule, designated by the symbol ~, are so-cal'

high energy phosphate bonds. Each of these bonds stores about 11,000 calories of energy per mole of ATP (under the physical conditions in the body). Therefore, when one phosphate radical is removed from the molecule, 11,000 calories of energy that can be used to energize the muscle contractile process are released. Then, when the second phosphate radical is removed, still another 11,000 calories become available. Removal of the first phosphate converts the ATP into *adenosine diphosphate* (ADP), and removal of the second converts this ADP into *adenosine monophosphate* (AMP).

Unfortunately, the amount of ATP present in the muscles, even in the well-trained athlete, is sufficient to sustain maximal muscle power for only 5 or 6 seconds, maybe enough for a 50 meter dash. Therefore, except for a few seconds at a time, it is essential that new adenosine triphosphate be formed continuously, even during the performance of athletic events. Figure 84–1 illustrates the overall metabolic system, showing the breakdown of ATP first to ADP and then to AMP, with the release of energy to the muscles for contraction. To the left-hand side of the figure are illustrated the three different metabolic mechanisms that are responsible for reconstituting a continuous supply of adenosine triphosphate in the muscle fibers. These are the following.

Release of Energy from Phosphocreatine. Phosphocreatine is another chemical compound that has a high energy phosphate bond, with the following formula:

$$\text{Creatine}\sim PO_3$$

This can decompose to *creatine* and *phosphate ion*, as illustrated to the left in Figure 84–1, and in doing so releases large amounts of energy. In fact, the high energy phosphate bond of phosphocreatine has slightly more energy than the bond of ATP. Therefore, the phosphocreatine can easily provide enough energy to reconstitute the high energy bonds of the ATP. Furthermore, most muscle cells have two to three times as much phosphocreatine as ATP.

A special characteristic of energy transfer from phosphocreatine to ATP is that it occurs within a small fraction of a second. Therefore, in effect, all the energy stored in the muscle phosphocreatine is instantaneously available for muscle contraction, just as is the energy stored in ATP.

The cell phosphocreatine plus its ATP are called the *phosphagen energy system*. These together can provide maximal muscle power for a period of 10 to 15 seconds, barely enough for the 100 meter run. Thus, the energy from the phosphagen system is used for maximal short bursts of muscle power.

The Glycogen–Lactic Acid System

The stored glycogen in muscle can be split into glucose and the glucose then utilized for energy. The initial stage of this process, called *glycolysis*, occurs entirely without use of oxygen and therefore is said to be *anaerobic metabolism* (see Chapter 67). During glycolysis, each glucose molecule is split into two *pyruvic acid molecules*, and energy is released to form several ATP molecules. Ordinarily the pyruvic acid then enters the mitochondria of the muscle cells and reacts with oxygen to form still many more ATP molecules. However, when there is insufficient oxygen for this second stage (the oxidative stage) of glucose metabolism to occur, most of the pyruvic acid is converted into *lactic acid*, which t diffuses out of the muscle cells into the interstitial f and blood. Therefore, in effect, much of the mu glycogen becomes lactic acid, but in doing so conside ble amounts of adenosine triphosphate are formed tirely without the consumption of oxygen.

Another characteristic of the glycogen–lactic acid s tem is that it can form ATP molecules about 2.5 ti as rapidly as can the oxidative mechanism of the m chondria. Therefore, when large amounts of adenos triphosphate are required for moderate periods of mus contraction, this anaerobic glycolysis mechanism can used as a rapid source of energy. It is not as rapid the phosphagen system, but about half as rapid.

Under optimal conditions the glycogen–lactic a system can provide 30 to 40 seconds of maximal mus activity in addition to the 10 to 15 seconds provided by the phosphagen system.

The Aerobic System

The aerobic system means the oxidation of foodstuffs in the mitochondria to provide energy. That is, as illustrated to the left in Figure 84–1, glucose, fatty acids, and amino acids from the foods—after some intermediate processing—combine with oxygen to release tremendous amounts of energy that are used to convert AMP and ADP into ATP, as was discussed in Chapter 67.

In comparing this aerobic mechanism of energy supply with the glycogen–lactic acid system and the phosphagen system, the relative *maximum rates of power generation* in terms of ATP utilization are the following:

Aerobic system	1 *M* of ATP/min
Glycogen–lactic acid system	2.5 *M* of ATP/min
Phosphagen system	4 *M* of ATP/min

On the other hand, when comparing the systems for endurance, the relative values are the following:

Prosphagen system	10 to 15 seconds
Glycogen–lactic acid system	30 to 40 seconds
Aerobic system	unlimited time (as long as nutrients last)

Thus, one can readily see that the phosphagen system is the one utilized by the muscle for power surges, and the aerobic system is required for prolonged athletic activity. In between is the glycogen–lactic acid system, which is especially important for giving extra power during such intermediate races as the 200 to 800 meter runs.

What Types of Sports Utilize Which Energy Systems?

By considering the vigor of a sports activity and its duration, one can estimate very closely which of the energy systems are used for each activity. Various approximations are presented in Table 84–1.

Recovery of the Muscle Metabolic Systems After Exercise

Performance in athletic events is often determined by how rapidly the athlete can recover strength between

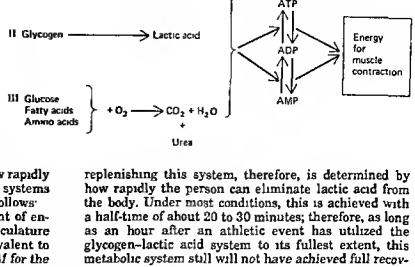


Figure 84–1. The three important metabolic systems that supply energy for muscle contraction.

surges of activity, and in general this means how rapidly the energy systems can recover. Each of these systems has its own characteristic rate of recovery, as follows:

The Phosphagen System. The total amount of energy in the phosphagen system in all the musculature of the well-trained male athlete's body is equivalent to about 0.6 mole per liter (*M*) of ATP (about 0.3 *M* for the female), and this can be almost completely depleted in an average of 10 to 15 seconds of maximal muscle activity. However, the glycogen–lactic acid system can replenish this phosphagen system as rapidly as 2.5 *M* of ATP per minute, and the aerobic system can replenish it as rapidly as 1 *M* per minute. Therefore, in theory, it would be possible for these other energy systems to replenish fully the phosphagen system within 15 to 30 seconds after its full depletion, which would mean that a person could run a second 100 meter dash in less than 1 minute after the first. However, in practice, this does not work quite that way because the other systems function at full *force* to replenish the phosphagen only when the phosphagen system is almost totally depleted. Instead, the phosphagen normally is replenished with a *half-time* of about 20 to 30 seconds. This means that for those events using the phosphagen system alone, such as high jumping, one could reasonably expect to have full replenishment of this system within about 3 to 5 minutes.

The Glycogen–Lactic Acid System. The limitation in the use of this system for energy is mainly the amount of lactic acid that the person can tolerate in muscles and body fluids. Lactic acid causes extreme *fatigue*, which serves as a self limitation to further use of this system for energy. The amount of time required for replenishing this system, therefore, is determined by how rapidly the person can eliminate lactic acid from the body. Under most conditions, this is achieved with a half-time of about 20 to 30 minutes; therefore, as long as an hour after an athletic event has utilized the glycogen–lactic acid system to its fullest extent, this metabolic system still will not have achieved full recovery.

The Aerobic System—Short-Term Recovery and the "Oxygen Debt." Recovery of the aerobic system has a short-term phase and a long-term phase, one lasting about an hour and the other several days. The short-term phase of recovery is a function of the so-called *oxygen debt*, illustrated in Figure 84–2. The oxygen debt is defined as the extra amount of oxygen that must be taken into the body after an athletic event to restore all the metabolic systems back to their full normal state. Oxygen debt can be accumulated in two different ways. First, part of this debt results from usage of oxygen that is already stored in different parts of the body. For instance, even normally about 0.3 liter of oxygen is stored in the muscles themselves combined with myoglobin, an oxygen-binding chemical substance similar to hemoglobin that is present in muscle fibers. In addition, almost 1 liter of oxygen is normally combined with all the hemoglobin in the blood, and another 0.5 liter is in the air of the lungs as well as about 0.25 liter dissolved in all the body fluids. Most of this oxygen can be used by the muscles during exercise and therefore must be replenished after the exercise is over.

Second, oxygen debt can be accumulated by depletion of both the phosphagen and the glycogen–lactic acid systems. As much as 2 liters of oxygen are required to

Table 84–1 ENERGY SYSTEMS USED IN VARIOUS SPORTS

Phosphagen system, almost entirely	
100 meter dash	Diving
Jumping	Football dashes
Weight lifting	
Phosphagen and glycogen–lactic acid systems	
200 meter dash	Baseball home run
Basketball	Ice hockey dashes
Glycogen–lactic acid system, mainly	
400 meter dash	Tennis
100 meter swim	Soccer
Glycogen–lactic acid and aerobic systems	
800 meter dash	2000 meter rowing
200 meter swim	1500 meter run
1500 meter skating	1 mile run
Boxing	400 meter swim
Aerobic system	
10,000 meter skating	Marathon run (26.2 mi, 42.2 km)
Cross-country skiing	Jogging

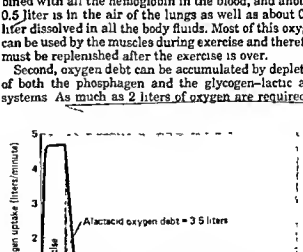


Figure 84–2. Rate of oxygen uptake by the lungs during maximal exercise for 4 minutes and then for almost 1 hour after the exercise is over. This figure demonstrates the principle of *oxygen debt*.

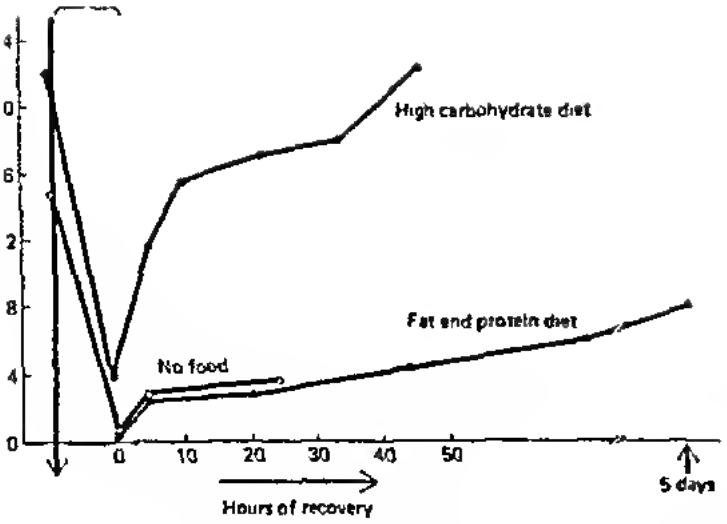


Figure 84–3. Effect of diet on the rate of muscle glycogen replenishment following prolonged exercise (Reprinted from Fox *Sports Physiology* Philadelphia, Saunders College Publishing, 1979)

nish a fully depleted phosphagen system, and as as 8 liters to replenish a depleted glycogen–lactic ystem

foto, a person can develop an oxygen debt up to 10 liters, and this oxygen is repaid to the body for as as an hour or more after periods of exhaustive ise Figure 84–2 illustrates that the repayment s in two different stages First, that portion of the not related to the accumulation of lactic acid, called *acid oxygen debt*, which means the oxygen needed plenish the oxygen stores of the body as well as to nish the phosphagen system, is usually fully repaid n 2 to 3 minutes. On the other hand, removal of actic acid from all the body fluids requires an hour ore, so that the *lactic acid oxygen debt*, which is lly by far the greater proportion of the total debt, nues to be repaid very slowly for at least an hour efore, for those sports that deplete the glycogen– acid metabolic system, for full recovery one should at least an hour and preferably two hours between ts

e mechanism of recovery of the glycogen–lactic acid m is simply to remove the lactic acid from the blood other body fluids This is achieved in two ways , some of the lactic acid is converted back into vic acid and then metabolized directly by all the tissues Second, much of the lactic acid is converted glucose by the liver, and the glucose in turn is used ly to replenish the glycogen stores of the muscles

e Aerobic System—Long-Term Recovery: Im-ance of Muscle Glycogen Stores. Earlier in the ter we discussed the importance of stored muscle ogen for muscle endurance. This is true because ogen is the food substrate of choice not only for the ogen–lactic acid system but also for the aerobic ative energy system as well The muscle endurance can be achieved may be as long as 4 hours of nustive exercise in the athlete who has a high entration of muscle glycogen, or as little as 1 5 s in the athlete with minimal muscle glycogen

covery from exhaustive muscle glycogen depletion ot a simple matter, requiring hours to days rather the seconds or minutes required for the phosphagen glycogen–lactic acid metabolic systems Figure illustrates this recovery process under three different conditions first, in persons on a high carbohydrate diet, second, in persons on a high fat/high protein diet and, third, in persons with no food. Note that on the high carbohydrate diet, full recovery occurred in approximately 2 days, whereas persons on the high fat/high protein diet or on no food at all showed extremely little recovery even after periods as long as days The message of this study is that it is important for an athlete, even when on a high carbohydrate diet not to participate in exhaustive exercise during the last 24 to 48 hours prior to a grueling athletic event.

Nutrients Used During Muscle Activity

Though we have emphasized the importance of a high carbohydrate diet and large stores of muscle glycogen for maximal athletic performance, this does not mean that only carbohydrates are used for muscle energy—it means simply that carbohydrates are used by preference Actually, the muscles often use large amounts of fat for energy in the form of *fatty acids* and *acetoacetic acid* (see Chapter 68) and also use to a much less extent proteins in the form of amino acids In fact, even under the best conditions, in those endurance athletic events that last longer than 4 to 5 hours the glycogen stores of the muscle become depleted and are then of little further use for energizing muscle contraction. Instead, the muscle now depends upon glucose that can be absorbed from the gut, which is limited, or upon energy from other sources, mainly from fats

Figure 84–4 illustrates the approximate relative usage of carbohydrates and fat for energy during prolonged exhaustive exercise under three different dietary conditions high carbohydrate diet, mixed diet, and high fat diet. Note that most of the energy is derived from carbohydrate during the first few seconds or minutes of the exercise, but at the time of exhaustion, as much as 50 to 80 per cent of the energy is being derived from fats rather than carbohydrates.

Not all the energy from carbohydrates comes from the stored muscle glycogen. In many persons almost as much glycogen is stored in the liver as in the muscles, and this can be released into the blood in the form of glucose then taken up by the muscles as an energy source In

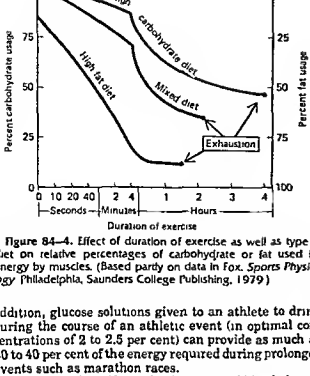


Figure 84–4. Effect of duration of exercise as well as type of diet on relative percentages of carbohydrate or fat used for energy by muscles. (Based partly on data in Fox, *Sports Physiology* Philadelphia, Saunders College Publishing, 1979)

addition, glucose solutions given to an athlete to drink during the course of an athletic event (in optimal concentrations of 2 to 2.5 per cent) can provide as much as 30 to 40 per cent of the energy required during prolonged events such as marathon races.

In essence, then, if muscle glycogen and blood glucose are available, these are the energy nutrients of choice for intense muscle activity Yet, even so, for a real endurance event one can expect fat to supply more than 50 per cent of the required energy after about the first 3 to 4 hours

Effect of Athletic Training on Muscles and Muscle Performance

Importance of Resistance Training. One of the cardinal principles of muscle development during athletic training is the following Muscles that function under no load, even if they are exercised for hours upon end, increase little in strength At the other extreme, muscles that contract at or near their maximal force of contraction will develop strength very rapidly even if the contractions are performed only a few times each day. Utilizing this principle, experiments on muscle building have shown that 6 maximal or nearly maximal muscle contractions performed in three separate sets 3 days a week give approximately optimal increase in muscle strength and without producing chronic muscle fatigue. The upper curve in Figure 84–5 illustrates the approximate percentage increase in strength that can be achieved in the previously untrained person by this optimal resistive training program, showing that the muscle strength increases about 30 per cent during the first 6 to 8 weeks but reaches a plateau after that time Along with this increase in strength is approximately an equal percentage increase in muscle mass, which is called *muscle hypertrophy*

Muscle Hypertrophy. The basic size of a person's muscles is determined mainly by heredity plus the level of testosterone secretion, which, in men, causes considerably larger muscles than in women. However, with training, the muscles can be hypertrophied perhaps an additional 30 to 60 per cent. Most of this hypertrophy results from increased diameter of the muscle fibers, but this is not entirely true because greatly enlarged muscle fibers can split down the middle along their entire length to form entirely new fibers, thus increasing the numbers of fibers as well

The changes that occur inside the hypertrophied muscle fibers themselves include (1) increased numbers of myofibrils, proportionate to the degree of hypertrophy; (2) increased numbers and sizes of mitochondria; (3) as much as 25 to 40 per cent increase in the components of the phosphagen metabolic system, including both ATP and phosphocreatine, (4) as much as 100 per cent increase in stored glycogen, and (5) as much as 75 to 100 per cent increase in stored triglyceride (fat). In addition, the enzymes required for the oxidative metabolic system are increased, increasing the maximum oxidation rate and efficiency of the oxidative metabolic system as much as 45 per cent.

Fast Twitch and Slow Twitch Muscle Fibers

In the human being, all muscles have varying percentages of *fast twitch* and *slow twitch muscle fibers*. For instance, the gastrocnemius muscle has a higher preponderance of fast twitch fibers, which gives it the capability of very forceful and rapid contraction of the type used in jumping On the other hand, the soleus muscle has a higher preponderance of slow twitch muscle fibers and therefore is said to be the muscle that is used to a greater extent for prolonged lower leg muscle activity.

The basic differences between the fast twitch and the slow twitch fibers are the following.

1 Fast twitch fibers are about two times as large in diameter.

2. The enzymes that promote rapid release of energy from the phosphagen and glycogen–lactic acid energy systems are two to three times as active in fast twitch fibers as in slow twitch fibers, thus making the maximal power that can be achieved by fast twitch fibers as great as two times that of slow twitch fibers

3. Slow twitch fibers are mainly organized for endurance, especially for generation of aerobic energy. They have far more mitochondria than the fast twitch fibers. In addition, they contain considerably more myoglobin, a hemoglobin-like protein that combines with oxygen within the muscle fiber; and even more important, myoglobin increases the rate of diffusion of oxygen

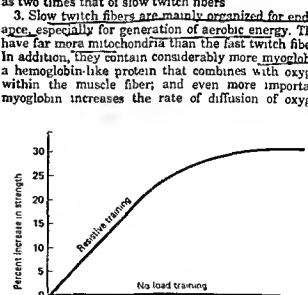


Figure 84–5. Approximate effect of optimal resistive exercise training on increase in muscle strength over a training period of 10 weeks.

ıout the fiber by shuttling oxygen from one mol- : myoglobin to the next. In addition, the enzymes ıerobic metobolic system are considerably more n slow twitch fibers than in fast twitch fibers ıe number of capillaries per mass of fibers is ın the vicinity of slow twitch fibers than in the ' of fast twitch fibers.

ımmary, fast twitch fibers can deliver extreme ıs of power for short periods of time. On the ather ılow twitch fibers provide endurance, delivering ;ed strength of contraction over much longer of time

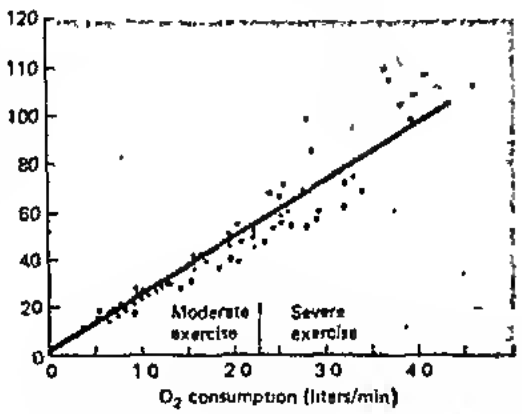
·ditory Differences Among Athletes for Fast ı Versus Slow Twitch Muscle Fibers. Some ı have considerably mora fast twitch than slaw fibers, and others have more slow twitch fibers; viously could determine to some extent the ath- pabilities of different individuals Unfortunately, : training has not been shown to change the : proportions of fast twitch and slow twitch fibers ır much an athlete mught wish to develop one ' athletic prowess over another. Instead, this Is ıned almost entirely by genetic inheritance, and turn helps determine which area of athletics Is ılted to each person some people are born to be ıoners, others are born to be sprinters and jump- r example, the following are recorded percentages twitch versus slow twitch fiber in the quadriceps s of different types of athletes·

	Fast Twitch	*Slow Twitch*
rathoners	18	82
ımmers	26	74
erage man	55	45
ıght lifters	55	45
ınters	63	37
npers	63	37

IRATION IN EXERCISE

gh one's respiratory ability ıs of relatively little ı for the performance of sprint types of athletics, ıtical for maximal performance in endurance ıs Let us see how important it Is

ɡen Consumption and Pulmonary Ventilation reise. Normal oxygen consumption for a young ıan at rest ıs about 250 ml per minute However,



re 84–6. Effect of exercise on oxygen consumption and ıory rate (From Gray Pulmonary Ventilation and its Phys- al Regulation. Springfield, Ill., Charles C Thomas.)

under maximal conditions this can bo increased to approximately the following average levels:

Untrained average male	3600 ml/min
Athletically trained average male	4000 ml/min
Male marathon runners	5100 ml/min

Figure 84–6 illustrates the relationship between oxygen consumption at different degrees of exercise and *total pulmonary ventilation*. It is clear from this figure, as would be expected, that there is a linear relationship. In round numbers, both oxygen consumption and total pulmanary ventilation increase about 20-fold between the resting state and maximum intensity of exercise.

The Limits of Pulmonary Ventilation. How severely do we stress our respirotory systems during exercise? This can be answered by the following comparison for the normal male.

Pulmonary ventilation at maximal exercise	100 to 110 liter/mia
Maximal breathing capacity	150 to 170 liter/min

Thus, the maximal breathing capacity is about 50 per cent greater than the actual pulmonary ventilation during maximal exercise. This obviously provides an element of safety for athletes, giving them extra ventilation that can be called on in such conditions as (1) exercise at high altitudes, (2) exercise under very hot conditians, and (3) abnormalities in the respirstory system.

The important point is that the respirstory system is not normally the most limiting factor in the delivery of oxygen to the muscles during maximal muscle aeroble metabolism. We shall see shortly that the ability of the heart to pump blood to the muscles is a greater limiting factor.

Effect of Training on Vo_2 Max. The abbreviation for the rate af oxygen usage under maximal aerobic metabolism Is Vo_2 Max Figure 84–7 illustrates the progressive effect of athletic training on Vo_2 Max recorded in a group of subjects beginning at the level of no training and then pursuing the training program for 7 to 13 weeks. In this study, it is surprising that the Vo_2 Max increased only about 10 per cent. Furthermore, the frequency of training, whether two times or five times per week, made little difference on the increase in Vo_2 Max Yet, as was pointed aut earlier, the Vo_2 Max al marathoners is about 45 per cent greater than that of the untrained person. Part of this greater Vo_2 Max af the marathoner is genetically determined, that is, those persons who have greater chest aizes and stronger respiratory muscles select themselves to become marathoners. However, it is also very likely that the very prolonged training of the marathoner does increase the Vo_2 Max by values considerably greater than the 10 per cent that has been recorded in short-term experiments such as that in Figure 84–7.

The O_2 Diffusing Capacity of Athletes. The O_2 *diffusing capacity* is a measuro af the rate at which oxygen can diffuse from the alveoli into the blood This is expressed in terms of *milliliters of oxygen that will diffuse for each millimeter of mercury difference between alveolar partial pressure of oxygen and pulmonary blood oxygen pressure* That is, if the partial pressure of oxygen in the alveoli is 91 mm Hg while the pressure in the blood is 90 mm Hg, the amount of axygen that diffuses through the respiratory membrane each minute is the diffusing capacity. The following are measured values for different diffusing capacities:

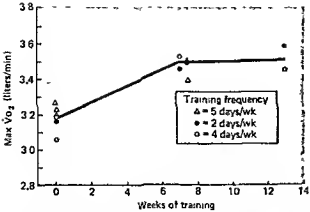


Figure 84–7. Increase in V_{O_2} Max over a period of 7 to 13 weeks of athletic training. (Reprinted from Fox. *Sports Physiology* Philadelphia, Saunders College Publishing. 1979)

Nonathlete at rest	23 ml/min
Nonathlete during maximum exercise	48 ml/min
Speed skaters during maximum exercise	64 ml/min
Swimmera during maximum exercise	71 ml/min
Oarsmen during maximum exercise	80 ml/min

The most startling fact about these results is the almost threefold increase in diffusing capacity between the resting state and the state of maximum exercise This results mainly from the fact that blood flow through many of the pulmonary capillaries is very sluggish or even dormant in the resting state, whereas in exercise increased blood flow through the lungs causes all the pulmonary capillaries to be perfused at their maximum level, thus providing far greater surface area through which oxygen can diffuse into the pulmonary capillary blood.

It is also clear from these values that those athletes who require greater amounts of oxygen per minute have higher diffusing capacities. Is this because persons with naturally greater diffusing capacities choose these types of sports, or is it because something about the training procedures Increases the diffusing capacity? The answer to this is not known, but one must believe that training does play some role In this, particularly the endurance ypes of training.

The Blood Gases During Exercise. Because of the reat usage of oxygen by the muscles in exercise, one vould expect the oxygen pressure of the arterial blood o decrease markedly during strenuous athletics and the arbon dioxide pressure of the venous blood to increase ar above normal. However, this is not the case Both hese remain nearly normal, illustrating the extreme ibility of the respiratory system to provide very adequate aeration of the blood even in heavy exercise. This llustrates another very important point. The blood gases do not have to become abnormal for respiration to be stimulated in exercise. Instead, respiration is stimulated mainly by neurogenic mechanisms. Part of this stimulation results from direct stimulation of the respiratory center by the same nervous signals that are transmitted from the brain to the muscles to cause the exercise. Part is believed to result from sensory signals transmitted into the respiratory center from the contracting muscles and moving joints. All this nervous stimulation of respiration is normally sufficient to provide almost exactly the proper increase in pulmonary ventilation to keep the blood respiratory gases—the oxygen and the carbon dioxide—almost normal.

Effect nf Smoking on Pulmonary Ventilation in Exercise. It is widely stated that smoking can decrease an athlete's "wind." This is a very true statement for many reasons. First, one effect of nicotine is to cause constriction of the terminal bronchioles of the lungs, which increases the resistance of air flow into and out of the lungs. Second, the irritating effects of smoke cause increased fluid secretion in the bronchial tree, as well as some swelling of the epithelial linings. Third, nicotine paralyzes the cilia on the surfaces of the respiratory epithelial cells that normally beat continuously to remove excess fluids and foreign particles from the respiratory tract. As a result, much debris accumulates in the respiratory passageways and adds further to the difficulty of breathing. Putting all these factors together, even the light smoker will feel respiratory strain during maximal exercise, and the level of performance obviously may be reduced

Much more severe are the effects of chronic smoking, because there is hardly any chronic smoker who does not eventually develop some degree of emphysema. In this disease, the following occur: (1) chronic bronchitis, (2) obstruction of many of the terminal bronchioles, and (3) destruction of many alveolar walls. In severe emphysema, as much as four fifths of the respiratory membrane can be destroyed; then even the slightest exercise can cause respiratory distress In fact, many such patients cannot even perform the athletic feat of walking across the floor of a single room without gasping for breath. Such is the indictment of smoking.

THE CARDIOVASCULAR SYSTEM IN EXERCISE

Muscle Blood Flow. The final common denominator of cardiovascular function In exercise is to deliver oxygen and other nutrients to the muscles For this purpose, the muscle blood flow increases drastically during exercise Figure 84–8 illustrates a recording of muscle blood flow in the leg calf of a person for a period of 6 minutes during strong intermittent contraction. Note the great increase in flow—about 13-fold—but note also that the flow decreased during each muscle contraction. Two points can be made from this study: (1) The actual contractile process itself temporarily decreases muscle blood flow because the contracting muscle compresses the intramuscular blood vessels; therefore, strong tonic

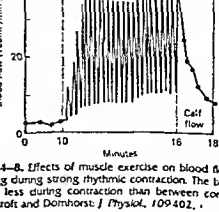


Figure 84–8. Effects of muscle exercise on blood flow in the calf of a leg during strong rhythmic contraction. The blood flow was much less during contraction than between contractions (from Barcroft and Dornhorst *J Physiol.* 109 402. .

contractions can cause rapid muscle fatigue because of lack of delivery of enough oxygen and nutrients during the continuous contraction. (2) The blood flow to muscles during exercise can increase markedly. The following comparison illustrates the maximum increase in blood flow that can occur in the well-trained athlete:

Resting blood flow	3 6 ml per 100 grams of muscle per minute
Blood flow during maximal exercise	90 ml per 100 grams of muscle per minute

Thus, muscle blood flow can increase a maximum of about 25-fold during the most strenuous exercise. About half this increase in flow results from intramuscular vasodilation caused by the direct effects of increased muscle metabolism, as was explained in Chapter 29 The other half results from multiple factors, the most important of which is probably the moderate increase in arterial blood pressure that occurs in exercise, usually about a 30 per cent increase The increase in pressure not only forces more blood through the blood vessels, but it also stretches the walls of the arterioles and further reduces the vascular resistance Therefore, a 30 per cent increase in blood pressure can often more than double the blood flow; this is in addition to the great increase in flow already caused by the metabolic vasodilation.

Work Output, Oxygen Consumption, and Cardiac Output During Exercise. Figure 84–9 illustrates the interrelationships among work output, oxygen consumption, and cardiac output during exercise. It is not surprising that all these are directly related to each other, as shown by the linear functions, because the muscle work output increases oxygen consumption, and oxygen consumption in turn dilates the muscle blood vessels, thus increasing venous return and cardiac output. Typical cardiac outputs at several levels of exercise are the following

Average young man at rest	5 5 liters/min
Maximum output during exercise in young untrained man	23 liters/min
Maximum output during exercise in male marathoner	30 liters/min

Thus, the normal untrained person can increase cardiac output a little over fourfold, and the well-trained

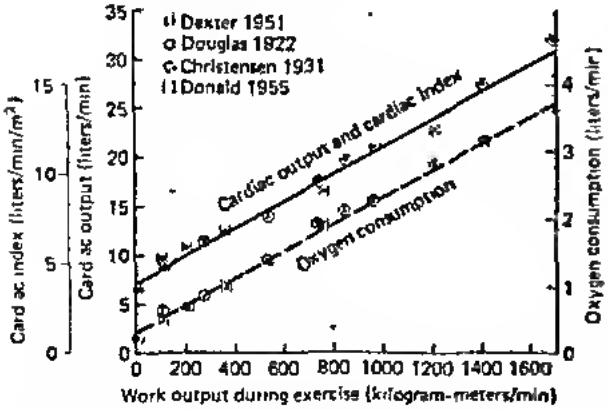


Figure 84–9. Relationship between cardiac output and work output (solid line) and betwen oxygen consumption and work output (dashed line) during different levels of exercise. (From Guyton, Jones and Coleman Circulatory Physiology Cardiac Output and Its Regulation. Philadelphia, W B Saunders Company, 3)

Table 84–2 COMPARISON OF CARDIAC OUTPUT BETWEEN MARATHONER AND NONATHLETE

		Stroke Volume (ml)	Heart Rate (beats/min)
Resting	Nonathlete	75	75
	Marathoner	105	50
Maximum	Nonathlete	110	195
	Marathoner	162	185

athlete can increase output about sixfold. Individual marathoners have been clocked at cardiac outputs as great as 35 to 40 liters per minute

Effect of Training on Heart Hypertrophy and on Cardiac Output. From the foregoing data, it is clear that marathoners can achieve maximum cardiac outputs about 40 per cent greater than those achieved by the untrained person. This results mainly from the fact that the heart chambers of marathoners enlarge about 40 per cent, and, along with enlargement of the chambers, the heart mass enlarges 40 per cent or more as well Therefore, it is not only the skeletal muscles but also the heart that hypertrophy during athletic training However, heart enlargement and increased pumping capacity occur only in the endurance types, not in the sprint types, of athletic training.

Even though the heart of the marathoner is considerably larger than that of the normal person, resting cardiac output is almost exactly the same as the normal However, this normal cardiac output is achieved by a large stroke volume at a reduced heart rate Table 84–2 compares stroke volume and heart rate in the untrained person and the marathoner

Thus, the heart-pumping effectiveness of each heart beat is 40 to 50 per cent greater in the highly trained athlete than in the untrained person, but there is a corresponding decrease in heart rate at rest.

Role of Stroke Volume and Heart Rate in Increasing the Cardiac Output. Figure 84–10 illustrates the approximate changes in stroke volume and heart rate as the cardiac output increases from its resting level of about 5 5 liters per minute to 30 liters per minute in the marathon runner The *stroke volume* increases from 105 ml to 162 ml, an increase of about 50 per cent, while the heart rate increases from 50 to 185 beats per minute, an increase of 270 per cent. Therefore, the heart rate increase accounts by far for a greater proportion of the increase in cardiac output than does the increase in stroke volume during strenuous exercise. The stroke volume reaches its maximum by the time the cardiac output has increased only half way to its maximum Any further increase in cardiac output must occur by increasing the heart rate.

Relationship of Cardiovascular Performance to Vo_2 Max. During maximal exercise, both the heart rate and the stroke volume are increased to about 95 per cent of their maximal levels. Since the cardiac output is equal to stroke volume *times* heart rate, one finds that the cardiac output is about 90 per cent of the maximum that the person can achieve. This is in contrast to about 65 per cent of maximum for pulmonary ventilation. Therefore, one can readily see that the cardiovascular system is normally much more limiting on Vo_2 Max than is the respiratory system For this reason, it is frequently stated that the performance that can be achieved by the marathoner mainly depends on his or her heart, for this is the most limiting link in the delivery of adequate oxygen to the exercising muscles.

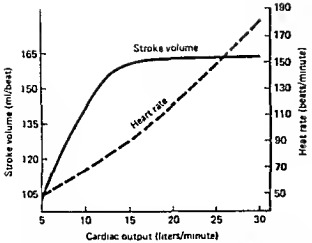


Figure 84–10. Approximate stroke volume output and heart rate at different levels of cardiac output in a marathon athlete.

Therefore, the 40 per cent advantage in maximum cardiac output that the marathoner has over the average untrained male is probably the most important physiological benefit of the marathoner's training program

Effect of Heart Disease and Old Age on Athletic Performance. Because of the critical limitation that the cardiovascular system places on maximal performance in endurance athletics, one can readily understand that any type of heart disease that reduces the maximum cardiac output will cause an almost corresponding decrease in achievable muscle power. Therefore, a person with congestive heart failure frequently has difficulty achieving even the muscle power required to climb out of bed, much less to walk across the floor

The maximum cardiac outputs of older persons also decrease considerably—as much as a 50 per cent decrease between the teens and the age of eighty years. Again, one finds that the maximum achievable muscle power is greatly reduced.

BODY HEAT IN EXERCISE

Almost all the energy released by the internal metabolism of nutrients is eventually converted into body heat. This even applies to the energy that causes muscle contraction, for the following reasons: First, the maximum efficiency for conversion of nutrient energy into muscle work, even under the best of conditions, is only 20 to 25 per cent; the remainder of the nutrient energy is converted into heat during the course of the intracellular chemical reactions. Second, almost all the energy that does go into creating muscle work still becomes body heat because all but a small portion of this energy is used for (1) overcoming viscous resistance to the movement of the muscles and joints, (2) overcoming the friction of the blood flowing through the blood vessels, and (3) other similar effects—all of which convert the muscle contractile energy into heat.

Now, recognizing that the oxygen consumption by the body can increase as much as 20-fold in the well-trained athlete and that the amount of heat liberated in the body is directly proportional to the oxygen consumption (as discussed in Chapter 71), one quickly realizes that tremendous amounts of heat are injected into the internal body tissues during endurance athletic events.

Next, coupling this vast rate of heat flow into the body with a very hot and humid day so that the sweating mechanism cannot eliminate the heat, the athlete can easily develop an intolerable and even lethal situation called *heat stroke*.

Heat Stroke. During endurance athletics even under normal environmental conditions the body temperature often rises from its normal level of 98.6°F to 102 to 103°F (37°C to 40°C). But, with very hot and humid conditions or great excesses of athletic clothing, the body temperature can then easily rise as high as 106 to 108°F. At this level the elevated temperature itself becomes destructive to tissue cells, especially destructive to brain cells. When this happens, multiple symptoms begin to appear, including extreme weakness, exhaustion, headache, dizziness, nausea, profuse sweating, confusion, staggering gait, collapse, and unconsciousness.

This whole complex is called "heat stroke," and failure to treat this immediately can lead to death In fact, even though the person has stopped exercising, the temperature does not easily decrease by itself. One of the reasons for this is that at these high temperatures the temperature-regulating mechanism itself often fails. A second reason is that the high temperature approximately doubles the rates of all intracellular chemical reactions, thus liberating still more heat

The treatment of heat stroke is to reduce the body temperature as rapidly as possible. The most practical way to do this is to remove all clothing, maintain a spray of water on all surfaces of the body or continually sponge the body, and blow air over the body with a strong fan. Experiments have shown that this can reduce the temperature either as rapidly or almost as rapidly as any other procedure, though some physicians prefer total immersion of the body in ice water containing a mush of crushed ice if this could possibly be available.

BODY FLUIDS AND SALT IN EXERCISE

As much as a 5 to 10 pound weight loss has been recorded in athletes in a period of 1 hour during endurance athletic events under hot and humid conditions Essentially all this weight loss results from loss of sweat. Loss of enough sweat to decrease body weight only 3 per cent can significantly diminish a person's performance, and a 5 to 10 per cent rapid decrease in weight in this way can often be very serious, leading to muscle cramps, nausea, and other effects Therefore, it is essential to replace fluid as it is lost.

Replacement of Salt and Potassium. Sweat contains a large amount of salt, for which reason it has long been stated that all athletes should take salt (sodium chloride) tablets when performing exercise on hot and humid days Unfortunately, overuse of salt tablets has led to more harm than good. Furthermore, if an athlete becomes acclimatized to the heat by progressive increase in athletic exposure over a period of 1 to 2 weeks rather than performing maximal athletic feats on the first day, the sweat glands also become acclimatized so that the amount of salt lost in the sweat is only a small fraction of that prior to acclimatization. This sweat gland acclimatization results mainly from increased aldosterone secretion by the adrenal cortex. The aldosterone in turn has a direct effect on the sweat glands to increase the reabsorption of sodium chloride

from the sweat before it issues forth onto the surface of the skin. Once the athlete is acclimatized, only rarely do salt supplements need to be considered during athletic events.

On the other hand, recent experience by military units suddenly exposed to heavy exercise in the desert has demonstrated still another electrolyte problem—the problem of potassium loss. This results partly from the fact that the increased secretion of aldosterone during heat acclimatization increases the loss of potassium in the urine as well as some increase in potassium in the sweat. As a consequence of these new findings, some of the newer supplemental fluids for athletics are beginning to contain properly proportioned amounts of potassium, usually in the form of fruit juices.

DRUGS AND ATHLETES

Without belaboring this issue, let us list some of the effects of drugs in athletics.

First, *caffeine* can increase athletic performance. In one experiment on a marathon runner, his running time for the marathon was reduced by 7 per cent by judicious use of caffeine in amounts similar to those found in a cup or so of coffee.

Second, use of *male sex hormones (androgens)* to increase muscle strength probably can increase athletic performance under some conditions, especially in women and in men who are poorly endowed with normal testosterone secretion. Unfortunately, some of the synthetic testosterone analog preparations can cause liver damage, and, in men, any type of male sex hormone preparation can lead to decreased testicular function, including both decreased formation of sperm and decreased secretion of the person's own natural testosterone. In a woman, even more dire effects can occur because she is not normally adapted to the male sex hormone.

Other drugs, such as *amphetamines* and *cocaine*, have been reputed to increase one's athletic performance. However, it is equally true that overuse of these drugs can lead to deterioration of performance. Furthermore, experiments have failed to prove the value of such drugs. Some athletes have been known to die during athletic performance because of interaction between such drugs and the norepinephrine and epinephrine released by the sympathetic nervous system during exercise. One of the causes of death under these conditions is overexcitability of the heart, leading to ventricular fibrillation, which is lethal within seconds.

REFERENCES

Adams, G. M. Exercise Physiology Primer. Costa Mesa, CA, Custom Publishing Co., 1983.

Alexander, R. M. Mechanics of skeleton and tendons. *In* Brooks, V. B. (ed.) Handbook of Physiology. Sec. 1, Vol. II. Bethesda, American Physiological Society, 1981, p. 17.

American Physiological Society. Skeletal Muscle. Baltimore, Waverly Press, 1983.

Appenzeller, O., and Atkinson, R. (eds.) Health Aspects of Endurance Training. New York, S. Karger, 1978.

Apple, D. F., Jr., and Cantwell, J. D.: Medicine for Sport. Chicago, Year Book Medical Publishers, 1979.

Armstrong, R. B. Mechanisms of exercise-induced delayed onset muscular soreness: A brief review. *Med. Sci. Sports Exerc.*, 16:529, 1984.

Baer, H. P., and Drummond, G. I. (eds.) Physiological and Regulatory Functions of Adenosine and Adenine Nucleotides. New York, Raven Press, 1979.

Basmajian, J. V.: Muscles Alive, Their Functions Revealed by Electromyography. Baltimore, Williams & Wilkins, 1978.

Blomqvist, C. G., and Saltin, B.: Cardiovascular adaptations to physical training. *Annu. Rev. Physiol.*, *45*:169, 1983.

Brengelmann, G. L. Circulatory adjustments to exercise and heat stress. *Annu. Rev. Physiol.*, *45*:191, 1983.

Bye, P. T. P., *et al.* Respiratory factors limiting exercise. *Annu. Rev. Physiol.*, *45*:439, 1983.

Dehn, M. M., *et al.* Clinical exercise performance. *Clin. Sports Med.*, *3*:319, 1984.

Dohm, G. L., *et al.* Protein metabolism during endurance exercise. *Fed. Proc.*, *44*:348, 1985.

Esmann, V. (ed.) Regulatory Mechanisms of Carbohydrate Metabolism. New York, Pergamon Press, 1978.

Fox, E. L. Sports Physiology. Philadelphia, W. B. Saunders Company, 1979.

Gollnick, P. D. Metabolism of substrates: Energy substrate metabolism during exercise and as modified by training. *Fed. Proc.*, *44*:353, 1985.

Guyton, A. C., *et al.* Circulatory Physiology: Cardiac Output and Its Regulation. 2nd Ed. Philadelphia, W. B. Saunders Company, 1973.

Hammond, H. K., and Froelicher, V. F.: Normal and abnormal heart rate responses to exercise. *Prog. Cardiovasc. Dis.*, *27*:271, 1985.

Harrison, M. H. Effects on thermal stress and exercise on blood volume in humans. *Physiol. Rev.*, *65*:149, 1985.

Hockey, R. Stress and Fatigue in Human Performance. New York, John Wiley & Sons, 1983.

Holloszy, J. O., and Booth, F. W. Biochemical adaptation to endurance exercise in muscle. *Annu. Rev. Physiol.*, *38*:273, 1976.

Homaher, F., and Keen, C. J. Skeletal muscle energetics and metabolism. *Annu. Rev. Physiol.*, *40*:93, 1978.

Killian, K. J., and Campbell, E. J. M. Dyspnea and exercise. *Annu. Rev. Physiol.*, *45*:465, 1983.

Klachko, D. M., *et al.* (eds.) Hormones and Energy Metabolism. New York, Plenum Press, 1978.

Lamb, D. R. Physiology of Exercise: Responses and Adaptations. New York, Macmillan Publishing Co., 1984.

Lind, A. R. Cardiovascular adjustments to isometric contractions: Static effort. *In* Shepherd, J. T., and Abboud, F. M. (eds.) Handbook of Physiology. Sec. 2, Vol. III. Bethesda, American Physiological Society, 1983, p. 947.

McManus: Cardiovascular Complications of Exercise. Chicago, Year Book Medical Publishers, 1985.

Mitchell, J. H., *et al.* The exercise pressor reflex: Its cardiovascular effects, afferent mechanisms, and central pathways. *Annu. Rev. Physiol.*, *45*:229, 1983.

Northrip, J. W., *et al.* Introduction to Biomechanic Analysis of Sport. Dubuque, Iowa, W. C. Brown, 1979.

Parmley, W. W., and Talbot, W. Heart as a pump. *In* Berne, R. M., *et al.* (eds.) Handbook of Physiology. Sec. 2, Vol. I. Baltimore, Williams & Wilkins, 1979, p. 429.

Partridge, L. D., and Benton, L. A. Muscle, the motor. *In* Brooks, V. B. (ed.) Handbook of Physiology. Sec. 1, Vol. II. Bethesda, American Physiological Society, 1981, p. 43.

Pearson, M. L. (ed.) Muscle Development: Molecular and Cellular Control. Cold Spring Harbor, NY, Cold Spring Harbor Institute, 1982.

Porter, R. (ed.) Human Muscle Fatigue. Woodstock, NY, Beekman, Publishers, 1981.

Rasch, P. J., and Burke, R. K. Kinesiology and Applied Anatomy: The Science of Human Movement. Philadelphia, Lea & Febiger, 1978.

Raab, J. Neuromuscular Atlas. New York, Praeger Publishers, 1984.

Rigotti, N. A., *et al.* Exercise and coronary heart disease. *Annu. Rev. Med.*, *34*:391, 1983.

Shepherd, J. T. Circulation to skeletal muscle. In Shepherd, J. T., and Abboud, F. M. (eds.) Handbook of Physiology. Sec. 2, Vol. III. Bethesda, American Physiological Society, 1983, p. 319.

Stone, H. L. Control of the coronary circulation during exercise. *Annu. Rev. Physiol.*, *45*:213, 1983.

Strauss, R. H. (ed.) Sports Medicine and Physiology. Philadelphia, W. B. Saunders Company, 1979.

Winter, D. A. Biomechanics of Human Movement. New York, John Wiley & Sons, 1979.

Index

A (anisotropic) bands, 121
a wave, 154, 228
Abdominal compression reflex, 252
Abdominal contraction, during labor, 993
  in cardiac output, 277
Abdominal recti, 466-467
Absorption, air, 525
  gastrointestinal See *Gastrointestinal absorption.*
  neonatal, 1002-1003
  tubular, 402-404
a-c interval, 229
Acapnia, 519
Acceleration, linear and angular, 623-825
Acceleratory forces, 532-534
Acclimatization, of sweating mechanism, 853
  to heat, 859
  to low oxygen pressure, 630-532
Accommodation, of cell membrane, 110
  of nerve fiber, 576
  visual, 704-705, 732
Accommodation reflex, 733
Acetate, in vascular control, 243
Acetazolamide, 449, 460, 775
Acetoacetate ions, reabsorption of, 404-405
Acetoacetic acid, formation in liver, *621-622, 837, 927-928*
Acetone, 928
Acetone breathing, 934-935
Acetylcholine, agonists and antagonists of, 138-139
  as neurotransmitter, 553
  as nicotinic drug, 696
  destruction of, 137
  formation of, 138, 551
  in action potential, 116
  in autonomic nervous system, 688-690
  in cardiac rhythmicity, 170
  in coronary blood flow, 298
  in myasthenia gravis, 139
  in neostriatum-substantia nigra system, 627
  in neuromuscular transmission, 137, 144
  in vasodilator system, 240
  local nature of, 876
  release of, 138
  secretion of, 137
Acetylcholine receptors, 689-690
Acetylcholine-gated ion channels, 91-92, 137-138
Acetylcholinergic drugs, for Parkinson's disease, 630
Acetylcholinesterase, 137, 551, 552, 689
Acetyl-coenzyme A (acetyl-CoA), in citric acid cycle, 23, 812
  in fatty acid metabolism, 820-821, *822, 837*
Acetyl-coenzyme A carboxylase (acetyl-CoA carboxylase), 823, 926
Achalasia, 762, 798
Achlorhydria, 775, 799
Acid, defined, 438
  fixed, 448
  metabolic, 448
  strong vs. weak, *438*
Acid hydrolases, 15
Acid base balance, clinical abnormalities of, 448-450
  neonatal, 1002
  renal regulation of, 444-448
  respiratory regulation of, 442-444
Acid base buffers, 439-442
Acidophils, 885, 891
Acidosis, 438-439
  chronic, ammonia buffer system in, 447
  diabetic, 928, 934
  effects on body, 449
  in renal failure, 455
  in shock, 330
  measurement of, 450
  metabolic, 448-449, 450, 458
  renal correction of, 445, 449-450
  respiratory, 448, 449-450
  synaptic transmission effect of, 561
  treatment of, 450
  uremic, 448
Acini, *770, 772, 923*
Acne, effect of testosterone on, 962
Acromegaly, 891-893
ACTH. See *Adrenocorticotropic hormone (ACTH).*
Actin, in ameboid locomotion, 24
  in mitosis, 37
Actin filaments, 121
  in muscle contraction, cardiac, 150
    skeletal, 124-126
    smooth, 140, 141
  molecular characteristics of, 124
Action potential(s), 106-110
  acute subthreshold, 116
  at axon hillock, 555-556
  biphasic, 118
  cardiac muscle, 151-152
    electrocardiogram of, 178
  elicitation of, 115-117
  monophasic, 118
    in cardiac muscle, 178
  neuronal somal potential in, 559
  plateaus in, 113
    in cardiac muscle, 151
Action potential(s) (*Continued*)
  in smooth muscle, 142
  propagation of, 110-111
  "recharging" after, 111-112
  recording of, 117-118
  rhythmic, 113-114
  skeletal muscle, 127
  smooth muscle, 140, 141-143, 144-145
  spike, 112
    in cardiac muscle, 151
    in smooth muscle, 142
  vs. receptor potential, 574
Action tremor, 645, 647
Activation gate, 107
Activator operator, 34
Activator protein, 34
Activator substance, 34
Active transport, 97-100
  ATP in, 841-842
  cellular, 19
  gastrointestinal absorption by, 792-795
  in resting membrane potential, 104-105
  insulin in, 928
  membrane potentials caused by, 103
  secondary, 794
  through tubular membrane, 400-401
  thyroid hormones in, 901
  vs. passive, 89
Adaptation, light and dark, 716-717
  of olfactory receptors, 749
  of sensory receptors, 575-576
  of taste, 748
  of thermal receptors, 604
Addisonian crisis, 920
Addison's disease, 427-428, 433, 435, 920
Adenine, in ATP, 23
  in DNA, 27, 28
  in RNA, 29
Adenohypophysis See *Anterior pituitary gland.*
Adenosine, as local vasodilator, 232-233, 296-297, 299
Adenosine diphosphate (ADP), control of glycolysis and oxidation by, 815
  formation of, 808
  in energy release, 23, 844
  in platelet plug, 77
  in rate of oxygen usage, 499
Adenosine monophosphate (AMP), complex with amino acid, 32
  formation of, 808
Adenosine triphosphate (ATP), 17
  cellular function of, 23-24, 841
  control of glycolysis and oxidation by, 815
  formation of, 22, 23

Arterial pressure load, and cardiac output, 158
Arterial pressure reflexes, 241
Arterial pressure regulation, 5
in cardiac output regulation, 274-275
in essential hypertension, 270
long term mechanisms for, 246
aldosterone in, 262-263
renal-body fluid system, 257-262
renin-angiotensin system, 262-264, 412-413
vs short-term mechanisms, 257
overall system for, 246-247
rapid mechanisms for, 246
hormonal, 253-255
intrinsic, 255
nervous, 247-253
vs long-term mechanisms, 257
Arterioles, afferent and efferent, 393, 394
blood flow through, 395
blood pressure in, 396
in glomerular filtration rate, 399-400, 412-413
characteristics of, 220
damping of pressure pulse in, 227-228
dilatation of, 242
edema from, 368
function of, 218
hepatic, 835
in microcirculation, 231, 348
resistance of, 219
terminol, 348
Arteriosclerosis, 226, 826
from hypertension, 270
in diabetes, 935
in hypothyroidism, 907
peripheral, 346
Arteriovenous anastomoses, 344, 850
Arteriovenous fistula, cardiac failure from, 310, 314-315
in cardiac output, 275, 278
Arteriovenous shunt See *Arteriovenous fistula*
Artery(ies), bronchial, 287
coronary See *Coronary artery*
distensibility of, 214
functions of, 218, 220
hepatic, 340, 341
pressure pulses in, 225-228
pulmonary, 287
umbilical, 985
uterine, 985
Articulation, 478, 662
Arytenoid cartilage, 478
Arytenoid muscles, 478
Ascites, from liver "sweating," 836
from portal obstruction, 342-343, 380
in thiamine deficiency, 868
Ascorbic acid, 870
Asparagine, 832
Aspartic acid, 832
Association areas, auditory, 653, 741-742
motor, 634
somatic sensory, 586-587, 653-654
thalamic, 653
visual, 653, 654, 727
Aster, 36
Astereognosis, 586
Asthenia, in diabetes, 934
Asthma, 68, 521
Astigmatism, 706-707
Astronauts See *Space physiology*
Ataxia, 647
Atelectasis, 294, 520-521
Atheromatous plaques, 826
Atherosclerosis, 826-827
ischemic heart disease from, 299-300
Atherosclerosis (*Continued*)
renal, 453
Athetosis, 629
Athletes See *Sports physiology*.
Athletic training. See *Exercise*.
ATP. See *Adenosine triphosphate (ATP)*
ATP synthetase, 23, 814
Atrial contraction, in electrocardiogram, 178
Atrial fibrillation, 174
electrocardiogram of, 202-203
in mitral valvular disease, 321
Atrial flutter, 172-173, 202
Atrial natriuretic factor, 427
Atrial paroxysmal tachycardia, 201-202
Atrial pressure, left, 288, 291
right, 221-224
in cardiac failure, 307, 308
in cardiac output and venous return curves, 282-284
Atrial pulsations, 228-229
Atrial reflexes, 250
Atrial sound, 156-157, 317
Atrial syncytium, 151
Atrioventricular block, 171
abnormal rhythms in, 198-199
complete, 199
partial, in atrial flutter, 172-173
Atrioventricular bundle, 151, 165, 168
premature contractions of, 200
Atrioventricular node, 165, 167-168
paroxysmal tachycardia of, 201-202
premature contractions of, 200
Atrioventricular valves, 155-156, 316-317
Atrium(ia), 150
as pumps, 153, 154
depolarization of, 187
in mitral valvular disease, 321
in respiratory unit, 486
pressures in, 154-155, 288, 291
repolarization of, 187-188
transmission of cardiac impulse through, 167
Atrophy, muscle, 134
Atropine, for bronchiolar constriction 475
Attention, generalized thalamocortical system in, 669
Attenuation reflex, 735
Audiogram, 744
Audiometer, 743-744
Auditory association areas, 653, 741-742
Auditory cortex, 653, 741-742
Auditory pathway, 740-741
Auditory radiation, 741
Auditory receptive aphasia, 662
Auditory relay nuclei, 741
Auerbach's plexus, 756-757
Auscultation, 317-318
Autoimmune diseases, 67
Autolysis, of cells, 20
Automaticity, of body, 7
Autonomic nervous system, 547. See also *Parasympathetic nervous system; Sympathetic nervous system*
circulatory regulation by, 237-240
coronary, 298
pulmonary, 289
control of accommodation by, 704, 732
control of gastrointestinal tract by, 757-758
control of heart by, 160-161, 170-171
control of pupillary aperture by, 732-733
control of sweating by, 852-853
Autonomic nervous system (*Continued*)
general organization of, 686-688
in homeostasis, 4
medullary, pontine, and mesencephalic control of, 695
pharmacology of, 696-697
psychosomatic effects transmitted by, 685
stimulation of gastrointestinal secretion by, 770-771
stimulation of insulin secretion by, 930
Autonomic reflexes, 694
for gastrointestinal activity. See *Gastrointestinal reflexes*.
in spinal cord, 617
Autoregulation, in hypertension, 265-266
of arterial pressure, 260-261
of blood flow, 234-235
in capillaries, 349
of glomerular filtration rate, 400, 410-413
Autoregulatory escape, 342
AV See *Atrioventricular*
Aversion, 680
Aviation physiology, accelerator forces in, 532-534
low oxygen pressure in, 529-530
temperature problems in, 534-535
Axon, 114, 550
Axon hillock, 555-556
Axon terminals. See *Terminal(s)*.
Axoneme, 26, 955
Axoplasm, 114
Azathioprine, 75
Azotemia, 455

B cells, 61-65
Babinski sign, 638
Bacteria, lysosomal digestion of, 54
Bainbridge reflex, 250
Baldness, 962
Ballistic movements, 645-646
Baroreceptor(s), 5, 247
adaptation of, 249
and body posture, 248
in blood volume control, 427
physiologic anatomy of, 247
response to pressure, 247-248
Baroreceptor nerves, hypertension from sectioning of, 269
Baroreceptor reflexes, 247-249, 694
in acute cardiac failure, 305
in compensated shock, 328
oscillation of, 253
Barometric pressure, 528
Barrel chest, 521
Basal channels, in renal tubules, 401
Basal conditions, 847
Basal ganglia, 626
damage to, 629-631
in limbic system, 676
motor functions of, 628-629
neuronal connections of, 626-628
Basal metabolic rate, 847-848
during pregnancy, 991
effect of estrogens on, 974
effect of testosterone on, 962
effect of thyroid hormone on, 901
Base(s), complementary, 30
DNA, 27, 28, 30
RNA, 29, 30
vs. acid, 438
Basement membrane, of glomerulus, 397

Basic electrical rhythm (BER), 755, 762-763
Basilar fibers, 736-737, 738-739
Basilar membrane, 735, 736-739
Basket cells, 643
Basolateral nuclei, 682
Basophils, 51, 58
heparin release by, 82
in infection, 65
Behavioral functions, amygdala in, 681-682
chemical transmitter systems for, 683-684
hippocampus in, 682-683
hypothalamus in, 677-681
limbic cortex in, 683
psychosomatic effects of, 681-685
Bends, 540
BER (basic electrical rhythm), 755, 762-763
Benheri, 868
blood flow in, 234
cardiac function in, 278, 310, 315
hypoxia in, 522
Beta cells, 923
Beta chains, 46
Beta globulins, 208
Beta oxidation, 820-821, 837
Beta receptors, 298, 337, 690
Beta waves, 669, 671
Beta-amino-isobutyric aciduria, 458
Beta cells, 635
Bicarbonate buffer system, 439-441
Bicarbonate ion(s), absorption of, 793, 796
concentration of, 436
in carbon dioxide transport, 501
in pancreatic juice, 779-780
in saliva, 772-773
ratio of chloride to, 448
tubular reabsorption of, 408, 445
Bicarbonate-chloride carrier protein, 501
Bile, bilirubin excretion in, 838-840
secretion of, 781-784
Bile canaliculi, 781, 835
Bile ducts, 781, 835
Bile salts, 782-783
emulsification of fat by, 788-789, 790
in absorption of fats, 795
Biliary pain, 600
Bilirubin, 48, 782
conjugated, 838, 839-840
excretion of, 838-840
free, 838, 839-840
Binocular vision, 708
Bipolar cells, 720, 721
Bitter taste, 745-746
Bladder, anatomy of, 460
atonic, 462
automatic, 462
autonomic control of, 694
emptying of, 461-462
innervation of, 460-461
tabetic, 462
uninhibited neurogenic, 462
Blastocyst, 984, 989
Bleeding, excessive, 83-84
Bleeding time, 85
Blind spot, 728
Blood, acetoacetic acid transport in, 821-822
acidity of, 502, 517
alkaline phosphatase in, 943
amino acids in, 829-831, 864, 915, 928-929
carbon dioxide concentration in, 507, 577
Blood (*Continued*)
carbon dioxide transport in, 500-502
chylomicron removal from, 818-819
cleansing of, by liver, 341
by spleen, 343
conductance of, 212
fat exchange between adipose tissue and, 820
glucose concentration in. See Blood glucose concentration.
hormone concentrations in, 878-879, 882-883
lipids in, effect of thyroid hormone on, 901
transport of, 818-819
neonatal, 1002
oxygen concentration of, 517
oxygen transport in, 494, 496-500
oxygen uptake by, 493-494
physical characteristics of, 206-208
shunted, 491, 494
sludging of, 329, 333
venous admixture of, 494
viscosity of, 207-208
Blood clot See also *Blood coagulation dissolution of*, 77
formation of, 77, 78-79
lysis of, 82-83
retraction of, 78-79
vicious cycle of, 79
Blood coagulation See also *Blood clot.*
disseminated intravascular, 84, 333
liver in, 838
mechanism of, 77-83
outside body, 85
prevention of, 82, 85
tests of, 85-86
Blood flow, 208-210 See also *Circulation*
blood pressure effect on, 213-214
cerebral, 338
in shock, 327
coronary, in ischemic heart disease, 299-303
in shock, 327
measurement of, 302-303
normal, 295-296
effect of thyroid hormone on, 902
effect on interstitial fluid oxygen pressure, 494-495
effect on metabolic use of oxygen, 499
in capillaries, 349-350
in different tissues and organs, 230-231
in exercise, *1015-1016*
in shock, 327, 330
kinetic energy of, 157
laminar, 209-210
pulmonary, 289-291
measurement of, 517
obstruction of, 293-294
resistance of, 208, 211-214
streamline, 209-210
through intestinal vessels, 342
through kidneys, 395-396, 399, 408
through liver, 340-342, 836
through placenta, 990
through skeletal muscles, 336-338
through skin, 343-345, 850
through spleen, 343
turbulent, 209-210
velocity of, 210, 219
Blood flow control, acute, 231, 232-234
"autoregulation," 234-235
by nervous system, 237-241
cerebral, 235, 338-340
coronary, 296-299
Blood flow control (*Continued*)
gastrointestinal, 342
humoral regulation, 241-243
in cardiac output regulation, 274
mechanisms of, 231-235
"metabolic," 234
"myogenic," 234-235
oxygen demand theory for, 233-234
reflexes for, 241
renal, 235, 413
through liver, 341
through skeletal muscles, 337
vasodilator theory for, 232
Blood gases, measurement of, 517
in respiratory control, 507-511
abnormalities of, 514-515
during exercise, 511-513, 1015
Blood glucose concentration, and glucagon, 931-932
effect of cortisol on, 914-915
effect of growth hormone on, 888, 889
effect on insulin secretion on, 929-930
fasting, 934
in diabetes, 933-934
in insulin shock, 936
regulation of, 932-933
Blood glucose-buffer system, 932
Blood groups, 70-74
Blood plasma See *Plasma.*
Blood poisoning, 326, 333
Blood pressure, 208 See also *Arterial pressure*
abdominal, 222
critical closing, 214
diastolic, 219, 225
measurement of, 245-246
effect on blood flow, 213-214
effect on vascular resistance, 213-214
high. See *Hypertension.*
in hepatic vessels, 838
intrarenal, 398
mean circulatory filling, 215
and blood reservoirs, 225
during exercise, 252
mean pulmonary filling, in left heart failure, 309
mean systemic filling, after cardiac failure, 306, 308
in cardiac output regulation, 278
in venous return, 281
measurement of, 210-211
reference level for, 224
systolic and diastolic, 245-246
pulmonary, 288, 291
renal, 396
standard units of, 210-211
static, 215
systolic, 219, 225
measurement of, 245-246
total tissue, 396
venous, 219, 221-224
yield, 214
Blood reservoirs, arteries as, 220
liver as, 341, 836
lungs as, 289
spleen as, 343
veins as, 224-225
Blood sinuses, in placental formation, 985
Blood transfusion See also *Transfusion reactions*
in shock, 334
Blood typing, 72, 73
Blood volume, 218, 383-384
aldosterone effect on, 912
and ADH secretion, 894
control of, 425-428

Blood volume (*Continued*)
determination of, 385
during pregnancy, 990-991
in aortic stenosis and aortic regurgitation, 320
in cardiac output regulation, 276, 278, 282, 426
in lungs, 288-289
in shock, 327
neonatal, 1001-1002
thyroid hormone effect on, 902
Blood brain barrier, 349, 377
Blood-cerebrospinal fluid barrier, 377
Blood-clotting factors, 79, 80-81, 83
Blue weakness, 718
Blue-sensitive pigment, 716
Body fluid compartments, 383-381
Body fluid volumes, changes in, 387-392
measurement of, 384-386
regulation of, 421-424
Body fluids See also *Extracellular fluid, Interstitial fluid, Intracellular fluid*
buffer systems of, 441-442
calcium and phosphate effects on, 940
conservation of, 421
in exercise, 1017-1019
in renal failure, 454-455
osmolality of, 388-389
osmotic equilibria in, 387-390
partition of, 383-384
Body temperature, and food intake, 864
deep, 854, 856-857
detection of, 853-851
during exercise, 1017
in premature infant, 1004
in respiratory control, 513
in shock, 333
mechanisms that decrease, 854
mechanisms that increase, 855-856
neonatal, 1003
normal, 949
regulation of, abnormalities of, 858-860
analysis of, 9-9
behavioral, 857
by hypothalamus, 241, 344, 679, 853-857
in extreme cold, 856, 859-860
insulator system in, 849
set point for, 856-857
Body water, intake vs output of, 382-383
regulation by hypothalamus, 679
total, 382, 386
Body weight, effect of thyroid hormone on 901-902
Bohr effect, 498, 986-987
Bone, absorption of, 942-943, 944-945
calcium in, 941-942, 944-945, 948
deposition of, 942, 943
diseases of 948-950
effect of estrogens on, 974
effect of testosterone on, 962
effect of vitamin D on, 946
fracture of 943
growth of, 889
phosphate in 941-942, 944-945
structure of, 941
Bone fluid, 945
Bone marrow, aplasia of, 48, 58
macrophages of, 55, 56
red blood cell formation by, 43
Bone salts, 941
Bone stress, 943
Bony labyrinth, 621
Botulinum toxin, and neuromuscular transmission, 138
and swallowing disorders, 798
Boutons See *Terminal(s)*
Bowman's capsule, 393
pressure in, 398
Boyle's law, 537
Bradycardia, 197
Bradykinin, 242, 342, 345, 521
Bradypnea, 516
Brain See also *Cerebral cortex.*
activation of, reticular system for, 665-668
thalamocortical system for, 668-669
behavioral functions of. See *Behavioral functions*
intellectual functions of See *Intellectual functions.*
Brain damage, from circulatory arrest, 335
neurogenic shock from, 332
Brain death, 667
Brain edema, 340, 514
Brain stem, connection of basal ganglia to, 627-628
control of subconscious, stereotyped movements by, 625-626
excitation of antigravity muscles by, 620-621
in pain sensation, 595-590
motor functions of, 619-626
motor pathways to spinal cord, 620
Brain tumor, headache from, 602
localization of, 670
Brain waves, 669-670
Braxton-Hicks contractions, 992
Breasts, development of, 994
effect of estrogens on, 974, 994
effect of progesterone on, 975, 994
Breathing See also *Respiration*
Cheyne-Stokes, 514-515
mouth-to-mouth, 478-479
onset at birth, 999-1000
periodic, 514-515
work of, 469
Broca's area, 634, 662
Brodmann area 17, 727
Bronchial slits, 475
Bronchioles, 474-475, 486
Bronchus(i), 474-475
Brown fat, 846, 855
Brown-Séquard syndrome, 601-602
Brunner's glands, 784
Brush border, in carbohydrate metabolism, 788
in digestion of peptides, 790
in reabsorption of protein, 405
in sodium transport, 400-401
Buccal glands, 772
"Buffalo" torso, 921
Buffer(s), acid base, 439-442
pK of, 440
Buffer nerves, 249
Buffer systems, acid base, 439
ammonia, 446-447
bicarbonate, 439-441
isohydric principle for, 442
of body fluids, 441-442
phosphate, 441, 446
protein, 441-442
quantitative dynamics of, 440-441
tubular, 446-447
Buffering power, 440-441
of respiratory system, 444
Bulboreticular facilitatory region, 612
Bulbourethral glands, 954
Bulk flow, of water through membrane channels, 96-97
through capillary membrane, 351
Bundle branch(es), 168
Bundle branch block, 171, 190-191
Bundle of His, 151, 165, 168
premature contractions of, 200
Bundle of Kent, 201-202
Burns, edema from, 369
shock from, 332
Bursa(e), 381
of Fabricius, lymphocyte processing in, 61, 62
in tolerance, 67
Bypass surgery, 303

C cells, 947
c wave, 154, 228
Cachexia, 866
Caffeine, and athletic performance, 1018
Caisson disease, 540
Calcarine fissure area, 727
Calcification, 941-942
metastatic, 949
Calcified plaques, 826
Calcitonin, 877, 946-948
Calcium, absorption of, 793, 937-939, 945-946
exchangeable, 942, 948
excretion of, 937, 945
fetal metabolism of, 998
for hypoparathyroidism, 949
metabolic functions of, 872
required by infant, 1003
Calcium channels, voltage-gated, 109-110
Calcium ion(s), activation of actin filament by, 121
in action potential, 109-110
in blood coagulation, 81, 85
in bone, 941-942, 944-945, 948
in cardiac failure, 308-309
in cardiac muscle contraction, 152-153, 162
in extracellular fluid, 428-433, 937-940, 944-946
in gastrointestinal muscle contraction, 756
in membrane excitability, 117
in neurotransmitter release, 551
in post tetanic facilitation, 560
in second messenger system, 881
in smooth muscle, 143-144
in vascular control, 243
released by sarcoplasmic reticulum, 127-129
Calcium ion concentration, and calcitonin secretion, 946-948
and parathyroid secretion, 946, 948
effect of testosterone on, 962
effect on vitamin D metabolism, 938-939
in extracellular fluid, 940, 944-946
in plasma, 939-940, 947-949
in rickets, 950
overall control of, 948
regulation of, 435-436
Calcium pulse, 129
Calcium pump, 98
in skeletal muscle, 129
in smooth muscle, 143
Calcium salts, 941-942
Calcium binding protein, 939
Calcium-sodium channels, 109-110
in cardiac muscle, 151-152, 166
in smooth muscle, 142
gastrointestinal, 756
in synaptic membrane, 551

Callus, 943
Calmodulin, 143, 881
Calorie, 844-845
Calorimetry, 845
Calsequestrin, 129
Canal of Schlemm, 378, 379
Cancer, 38-39
Capacitation, of sperm, 957
Capillary(ies), blood flow in, 349-350
  cannulation of, 353
  carbon dioxide diffusion to, 495-496
  characteristics of, 220
  function of, 218
  glomerular, colloid osmotic pressure in, 398-399
  in microcirculation, 231
  lymphatic, 361, 364
    obstruction of, 369, 380
  muscle, during exercise, 336
  oxygen diffusion from, distance of, 199
    to cells, 495
    to interstitial fluid, 494-495
  pentubular, 395, 396
  pulmonary, 288, 291-294
  resistance of, 219
  true, 231, 348
Capillary beds, 348-349
  glomerular, 395
  high pressure vs. low pressure, 395
  pentubular, 395-396
Capillary exchange, capillary and interstitial fluid pressure in, 357-358
  diffusion in, 3, 220-221, 350-351
  in lungs, 292-293, 372
  in potential spaces, 379-380
  Starling equilibrium for, 358-359
Capillary fluid shift, 255
Capillary membrane, diffusion through, 3, 220-221, 350-351
  fluid movement through, 352-359
  imbalances at, 359, 366
  pores in, 349
Capillary permeability, 242
  in edema, 369-370
  in potential spaces, 380
  in shock, 329
  relative 350-351
Capillary pores, 349, 350-351
Capillary pressure, average, 349
  functional, 354
  in edema, 368, 372-373
  in potential spaces, 380
  measurement of, 353-354
  peripheral, in cardiac failure, 311
  pulmonary, 288, 291, 292-293, 371-373
Capillary pulsation, 228
Capillary system, 348-349
  average function of, 349-350
Capillary wall, 349
Carbachol, 138
Carbaminohemoglobin, 501
Carbohydrate(s), absorption of, 794
  as protein sparers, 862
  composition of, 787
  content in different foods, 861
  depletion of, 866
  digestion of, 787-788
  fat-sparing effect of, 823
  formation of, 816-817
  in cell, 12
  in cell membrane, 14
  needed for growth hormone effects, 888
  triglyceride synthesis from, 822-823
Carbohydrate metabolism, 22
  blood glucose in, 817
  central role of glucose in, 809
  during exercise, 1012-1013
Carbohydrate metabolism (*Continued*)
  effect of cortisol on, 914
    in Cushing's syndrome, 921
  effect of growth hormone on, 888-889
  effect of insulin on, 924-926
  effect of thyroid hormone on, 901
  energy release in, by glycolytic pathway, 811-816
    by phosphogluconate pathway, 816
  formation of carbohydrates, 816-817
  glucose transport through cell membrane, 809-810
  liver in, 837
  rate of, 562-563
  "switching" between lipid metabolism and, 930
Carbon dioxide, at high altitudes, 528-529
  concentration of, in alveoli, 485, 490-492
    in blood, 507, 508
    in cerebrospinal fluid, 507
    in extracellular fluid, 444
  diffusion of, 495-496
    capacity for, 489
    through placental membrane, 987
  dissolved, 500, 501
  effect on oxygen hemoglobin dissociation curve, 498
  in acid base balance, 442
  in cerebral blood flow regulation, 339
  in vascular control, 235, 243
  regulation of, 5
  release of, by phosphogluconate pathway, 816
    in citric acid cycle, 812, 813
  removal of 4
  toxicity, in deep sea diving, 539
  transport of, 500-502
Carbon dioxide dissociation curve, 501
Carbon dioxide pressure, 493-496, 501
  determination of, 517
  in Cheyne-Stokes breathing, 514-515
  in deep sea diving, 539
  in entrapped air, 525
  in hypercapnia, 523
  in respiratory control, direct effect, 506-508
    during exercise, 511-513
  vs oxygen feedback control, 509-510, 511
Carbon dioxide-alveolar ventilation stimulation curve, 514
Carbon monoxide, combination of hemoglobin with, 500
  diffusing capacity for, 490
Carbon monoxide poisoning, 500, 518, 543
Carbon tetrachloride, 453
Carbonic acid, dissociation of, 440, 444, 501
Carbonic anhydrase, 42, 444, 501
Carbonic anhydrase inhibitors, 449, 460
Carboxypolypeptidase, 242, 778-779, 790
Carcinogens, 38
Carcinoid tumors, 242
Cardiac. See also *Heart*
Cardiac arrest, 175-176
  circulatory arrest from, 335
Cardiac arrhythmias, 171-176
  electrocardiographic interpretation of, 197-203
Cardiac contractility, 160-161
  assessment of, 163-164
Cardiac cycle, 153-157
  electrocardiogram of, 179-180
  volume-pressure diagram in, 159
Cardiac depression, in progressive shock, 328
Cardiac edema, 311-312, 368
Cardiac failure, acute, 305-307, 309, 310
  bilateral, 305
  cardiac reserve in, 312-313
  Cheyne-Stokes breathing caused by, 515
  chronic, 306-307, 310
  circulation in, 305-309
    pulmonary, 289, 291
  compensated, 305-306, 307, 313
  decompensated, 308-309, 314-315
  defined, 305
  edema in, 308, 311-312, 320, 368
  from hypertrophy of heart, 324
  from thiamine deficiency, 868
  graphical analysis of, 313-315
  high cardiac output, 310, 314-315
  low cardiac output, 273, 274, 300, 305, 310-312
  physiological classification of, 312
  severe, 307, 308-309
  unilateral, 305, 309-310, 320
Cardiac function, after myocardial infarction, 302
  regulation of, 157-164, 238, 239
Cardiac function curves, 161
Cardiac glands, 774
Cardiac impulse, transmission of, 167-169
Cardiac index, 272
Cardiac massage, 175
Cardiac muscle, action potentials in, 151-152
    electrocardiogram of, 178
  as syncytium, 150-151
  contraction of, 152-153
    and coronary blood flow, 295-296
    duration of, 153
    efficiency of, 157
    energy for, 157
    strength of, 160-161, 163-164
  hypertrophy of, 319-320, 324
  physiology of, 150-163
Cardiac output, after acute myocardial infarction, 300
  age and, 272
  aldosterone effect on, 912
  and arterial pressure load, 168, 246, 260
  and circulatory shock, 326, 327
  at high altitudes, 531
  blood volume and, 282
  calculation of, 157
  defined, 209, 272
  during pregnancy, 990
  high, 278
    in heart failure, 310, 314-315
  in exercise, 272
    analysis of, 283-284
    and pulmonary circulation, 290-291
    sympathetic nervous system in, 338
    with heart hypertrophy, 273, 1016-1017
  in hypertension, 265, 266, 267
  low, 277-278
  measurement of, 284-285
  metabolism effect on, 272
  normal values for, 272
  of newborn, 1002
  oxygen consumption and, 297
  skeletal nerves and muscles in, 252
  sympathetic nervous system and, 273, 282-283, 338
  thirst and, 431
  thyroid hormone effect on, 902
  veins in, 251-252

Cardiac output (*Continued*)
volume pressure diagram of, 159
Cardiac output curves, 159, 279-280, 282-284
after partial recovery from myocardial infarction, 307
in cardiac failure, 305, 308, 313-315
Cardiac output regulation, arterial pressure regulation in, 274-275
arteriovenous fistulae effect on, 275, 278
blood volume in, 276, 278, 282, 426
during exercise, 276-277, 338
extracardiac pressure in, 279-280
graphical analysis of, 278-284
heart in, 272-274, 276, 279
local blood flow regulation in, 274
local tissue metabolism in, 274-275
mean systemic filling pressure in, 276
special problems in, 275-277
total peripheral resistance in, 274-275
venous return in, 273
Cardiac pain, 600. See also *Angina pectoris*
Cardiac pumping, atrial fibrillation effect on, 174, 203
heart rate in, 158, 160
heart sounds and, 156-157
permissive level of, 272-274, 276
regulation of, 157-160
ventricular fibrillation effect on, 174
Cardiac reserve, after myocardial infarction, 302
defined, 312
diagnosis of low, 312-313
in compensated cardiac failure, 307
in patent ductus arteriosus, 323
Cardiac shock, 278, 300
Cardiac surgery, extracorporeal circulation in, 324
Cardiac tamponade, 280, 301
Cardiac toxicity, from hyperkalemia, 912
Cardiac veins, 295
damming of, 303
Cardiac work load, chronic increase in, 161, 270
Cardiac work output. See *Cardiac output*
Cardiogenic shock, 300, 310-311
Cardio-Green dye, 285
Cardiotachometer, 197-198
Cardiotonic drugs, 300
Cardiovascular system, deterioration in progressive shock, 328-330
effect of thyroid hormone on, 902
hypothalamic regulation of, 678-679
in exercise, 1015-1017
thiamine deficiency and, 868
Caries, 952-953
Carina, 475
Carnitine, 820
Carotenoid pigments, 867
Carotid bodies, 250, 508
Carotid sinus, 247
Carotid sinus syncope, 249
Carotid sinus syndrome, 197, 198, 249
Carrier proteins, 13-14
bicarbonate-chloride, 501
in active transport, 97
in facilitated diffusion, 92-93
of glucose, 810
in renal tubules, 401, 444
Carrier wave, 569, 741
Cartilage, growth of, 889
Castration, 963
Catalase, 16
Cataracts, 707
Catechol-O methyl transferase, 689
Cathodal current, 116
Cathode ray oscilloscope, 117-118
Caudate nucleus, 626-629
Celiac disease, 801
Celiac ganglion, 757
Cell(s), active transport in, 19
ameboid locomotion of, 24-25
animal, vs. precellular life forms, 18-19
autolysis of, 20
basic characteristics of, 2
carbohydrates in, 12
chromatin material in, 18
ciliary movement of, 25-26
cytoplasm of, 11, 14-18
deterioration, in shock, 330
digestion by, 20-21
energy systems of, 22-24, 843-844
enzyme activity in, 34-35
functional systems of, 19-26
genetic regulation of, 33-34, 37
growth of, 37
life cycle of, 35
lipids in, 12
nucleus of, 11, 18, 19
organelles of, 14-18, 19
organization of, 11-12
phagocytosis by, 19, 20
protein synthesis in, 21, 31-33
pseudopodium of, 24-25
reproduction of, 35-37
size of, regulation of, 37
structure of, 12-18
water in, 11
Cell differentiation, 33, 37-38
Cell membrane, 11. See also *Membrane potential(s)*
accommodation of, 110
as electrical capacitor, 101
carbohydrates in, 14
composition of, 12-14
diffusion through, 89-93
glucose transport through, 809-810
lipid barrier of, 12-13, 88
nuclear, 11, 18
osmosis across, 94-97
permeability of, 93, 95
proteins in, 13-14, 88-89
replenishment of, 22
Cell volume, Na^+-K^+ pump in control of, 98
Cellular sheets, active transport through, 99-100
Cementum, 951
Central lacteal, 791
Central nervous system (CNS), circulatory regulation by, 238-241
functional levels of, 548-549
in acidosis and alkalosis, 449
in homeostasis, 4
ischemic response in, 250-251
in shock, 328
oscillation of, 253
neuronal pools in, 563-569
pain signal transmission in, 594-596
smell transmission into, 750-751
sympathetic vasoconstrictor system control by, 239-240
synapses in. See *Synapse(s)*.
taste transmission into, 747-748
thiamine deficiency and, 867-868
thyroid hormone effect on, 902
vestibular apparatus and, 622-623
Central vein, 835
Central venous pressure, 221-224
Central venules, of liver, 340
Centrifugal acceleratory force, 532-533
Centrifugal nerve fibers, 566
Centrioles, 16, 36
Centromere, 36
Cephalins, 824
Cerebellar hemisphere, 639, 640
Cerebellar nystagmus, 647
Cerebellum, 638-639
anatomical functional areas of, 639-641
clinical abnormalities of, 647
damping function of, 645
in movement control, 643-647
neuronal circuitry of, 641-643
Cerebral circulation, 338-340
Cerebral cortex. See also *Brain; Motor cortex*
connections with basal ganglia, 627
functions controlled by, 548
in hearing, 741-742
in pain sensation, 596
intellectual functions of. See *Intellectual functions*
memory storage by, 548
physiologic anatomy of, 651-652
vasomotor center control by, 240
Cerebral edema, 532
Cerebral hemorrhage, from hypertension, 270
Cerebrospinal fluid (CSF), 374-377
carbon dioxide concentration in, 5[illegible]
Cerebrospinal fluid pressure, 378
headache from low, 602-603
Cerebrovascular disease, respiratory center depression in, 514
CGS (centimeter gram-second) units [illegible]
Chain terminating sequence, 30, 32
Cheilosis, 869
Chemical potential, 38[illegible]
Chemiosmotic mechanism, of mitochondria, 813-814
Chemoreceptor(s), 572
arterial pressure control by, 250
in respiratory control, 508-511
Chemoreceptor reflex, oscillation of [illegible]
Chemoreceptor trigger zone, [illegible]
Chemotaxis, by neutrophils and macrophages, 53, 56, 65
in ameboid locomotion, 25
positive vs. negative, 25
Chenodeoxycholic acid, 782, 78[illegible]
Chewing, 759-760
Cheyne-Stokes breathing, 514-5[illegible]
Chief cells, 774, 944
Child, growth and development [illegible]
Chills, 858-859
Chloride channels, in postsynaptic membrane, 551
Chloride ion(s), absorption of, [illegible]
alkalosis from loss of, 449
in action potential, 110
in inhibitory postsynaptic potential, 556
tubular reabsorption of, 405
Chloride ion concentration, across neuronal somal membrane, 5[illegible]
extracellular fluid, 436
plasma, 448
Chloride shift, 501
Chlorthiazide, 460
"Chokes," 510
Cholecalciferol, 871, 938
Cholecystokinin, 758, 765
as vasodilator, 342
chemical composition of, 777
inhibition of eating by, 864

lecystokinin (*Continued*)
hibition of gastric secretion by, 778
cal nature of, 876
gulation of bile release by, 782
gulation of pancreatic secretion by, 780-781
lecystokinin-8, 886
lera toxin, 784-785, 802
lesterol, 825-826
igestion of, 788
xogenous vs. endogenous, 825
n cell, 12
n cell membrane, 13
n chylomicrons, 818
secretion of, 782, 783-784
olesterol deposits, 826
olesterol ester(s), 788, 789-790
olesterol ester hydrolase, 790
olesterol esterase, 779
olic acid, 782, 825
oline, 825
oline acetyltransferase, 551
olinergic blockers, 696
olinergic fibers, in autonomic nervous system, 688-689
olinesterase, 137, 551, 552, 689
ondroitin sulfate, 21, 941
orda tympani, 747
ordae tendineae, 156
ordamesoderm, primordial, 38
orea, 629
orionic gonadotropin. See *Human chorionic gonadotropin.*
oroid, 713
oroid plexus, secretion by, 375
romatic aberration, 732
, 36
romatin roaterial, 18
romosomes, 18, 36
laughter, 37
livision of, 955
replication of, 36-37
lex, 955, 983
ironaxie, 117
) lomicrons, 795, 796, 818-819
yme, 762, 763, 766
neutralization of, 780
gmotrypsin, 778-779, 790
motrypsinogen, 779
, ry body, 378
ry ganglion, 732
78 y movement, 25-26
espiratory passageways, 475
-5y muscle, 690, 704
y nerves, 732
t processes, 378
(ia), 25-26
spectactory apparatus, 749
stibular apparatus, 621-622
79 line, 777, 800
ate *gyrus*, 676, 683
an rhythm, of glucocorticoid
otetretion, 919
r fibers, 690, 704
tion. See also *Blood flow*
ros al, 338-340
e, 59 eral, 236-237
heart, 299-300
ortal obstruction, 342
ary. See *Coronary circulation.*
eous, 343-345
g exercise, 337-338
orporeal, 324
er, 218
ral regulation of, 241-243
diac failure, 305-309

Circulation (*Continued*)
in valvular heart disease, 319-321
nervous regulation of. See *Sympathetic nervous system, circulatory regulation by.*
peripheral, 218
pulmonary. See *Pulmonary circulatory system.*
reflex regulation of, 241
splanchnic, 340-343
splenic, 343
storage areas of, 215
systemic, 206, 218-219
venous, cardiac, after acute myocardial infarction, 300-301
cross sectional area of, 219
volume-pressure curves in, 214-215
with congenital heart defects, 321-324
Circulatory arrest, 335
Circulatory capacity, and blood volume control, 427
Circulatory shock, causes of, 326-333
defined, 278, 326
effects of, 333-334
from aldosterone, 912
in diabetes, 934
in heat stroke, 859
renal ischemia caused by, 453
stages of, 326-327
treatment of, 334-335
Circulatory stress, shift of blood flow in, 341, 345
Circulatory system, as circuit, 206
at birth, 1000-1001
during pregnancy, 990-991
effects of anemia on, 49
effects of centrifugal accelerator force on, 533
effects of polycythemia on, 49-50
fetal, 997-998, 1000
in acclimatization to high altitude, 531
portal, 340-343
pulmonary. See *Pulmonary circulatory system*
transport of extracellular fluid by, 3-4
Circus movement, 172-175
Cirrhosis, of liver, 342, 836
Cisterna magna, 375
Citrate ion, as anticoagulant, 81, 85
in control of glycolysis, 815, 923
in fat synthesis and storage, 926, 928
in vascular control, 243
Citric acid cycle, 23, 812-813
Citrulline, 833
Clasmatocytes, 55
Clearance, free water, 417
metabolic, of hormone, 862-863
osmolar, 417, 422
para aminohippuric acid, 408
plasma, 407-408
tests of renal, 458-459
Climacteric, male, 964
Climate, effect on metabolic rate, 846
Climax, female, 980
Climbing fibers, 642
Clitoris, 980
Clonus, 612
Clot. See *Blood clot.*
Clotting factors, 79, 80-81, 83
Clotting time, 85-86
CNS See *Central nervous system (CNS).*
Coactivation, of motor neurons, 611
Coagulation. See *Blood coagulation.*
Coarctation, of aorta, 236, 321
Coated pits, 19
Cobalamin, 869

Cocaine, as membrane stabilizer, 117
in athletic performance, 1018
Cocarboxylase, 867
Cochlea, 734
conduction of sound to, 734-735
frequency determination by, 739
functional anatomy of, 735-737
loudness determination by, 739-740
organ of Corti in, 738-739
resonance in, 736-737
transmission of sound in, 737-738
Cochlear duct, 621
Cochlear nerve, 738
Cochlear nuclei, 740
Codons, 29, 30
Coenzyme A, 870. See also *Acetyl-coenzyme A (acetyl-CoA).*
Colchicine, 37
Cold. See also *Temperature.*
effect on skin circulation, 345
effect on TRH and TSH secretion, 904
exposure to extreme, 856, 859-860
in Raynaud's disease, 345-346
Cold receptors, 603-605, 854
Colitis, ulcerative, 768, 803
Collagen, 790
Collagen fibers, 941
Collagen monomers, 941
Collateral circulation, 236-237
in heart, 299-300
in portal obstruction, 342
Collecting duct, 394, 400
hydrogen ion secretion in, 444
osmolal concentration in, 417
potassium secretion by, 420
reabsorption by, 403-404
sodium transport through, 419-420
Collecting tubule, 394, 400
osmolal concentration in, 417
potassium reabsorption by, 420
potassium secretion by, 420
reabsorption by, 403, 419
sodium transport through, 419-420
Colloid, 897
Colloid osmotic pressure, albumin in, 208, 831
in glomerular capillaries, 398-399
interstitial fluid, 353, 357
plasma, 221, 352, 353, 356-357
in edema, 369, 380
in fluid volume excretion, 422-423
Colon. See also *Large intestine.*
absorption in, 796
irritable, 802, 803
movements of, 767-769
Colony-stimulating factor, 57
Color blindness, 718-720
Color vision, cones in, 716, 718
ganglion cells in, 723
nervous system interpretation in, 718
theories of, 718
visual cortex in, 726-727
Colostrum, 994
Coma, 667
diabetic, 928, 935
hepatic, 837
uremic, 455
Commissure of Probst, 741
Communication, 662
Compensation, for acute cardiac failure, 305-306, 307, 313
in control system, 8
in mitral valvular disease, 321
in shock, 326-328
Compensatory pause, 200
Competitive binding assays, 882

Complement, 53-54, 60, 64-65
Complex cells, 726
Compliance, delayed, 216, 368
  in interstitial spaces, 367
  pulmonary, 468-469
    and vital capacity, 471
    decreased, 518
  vascular, 214-216
    and pressure pulse, 225-226
Compliance work, 469
Compressed air sickness, 540
Concentration, mental, 656
Concentration difference, across neuronal somal membrane, 554
  effect on diffusion, 93-94, 481
    through capillary membrane, 351
  for water, 94
  in Nernst equation, 102
Condensation, 787
Conductance, of blood, 212-213
Conduction, decremental, 558-559
  electronic, 720
  heat loss by, 851-852
  in heart, 165-171
  one way, 549-550
  saltatory, 114-115
  velocity of, 115
    in cardiac muscle, 152
    in skeletal muscle, 127
Conduction deafness, 743, 744
Conductive membrane, 114
Cones, 711-712
  in color vision, 716, 718
  stimulation of, 720-721
Congenital anomalies, of heart, 321-324
Consciousness, 656-661
Consolidation, of memory 660, 683
Constant field equation, 102
Constipation, 802
  headache from, 603
Constrictor waves, 762-763
Contact junctions, 144
Contact lenses, 707
Contraception, 981
Contraction(s), abdominal, during labor, 993
    in cardiac output, 277
  after denervation, 134
  blood flow during, 336
  Braxton Hicks, 992
  cardiac muscle, 152-153
    atrial, electrocardiogram of, 178
    duration of, 153
    efficiency of, 157
    energy for, 157
    strength of 160-161, 163-164
    ventricular, electrocardiogram of, 178
  damping function in, 610
  duration of, heart rate and, 153
  efficiency of, 130
  energy for, 129-130, 841
  hunger 763, 863
  initiation of, 126-129
  isometric vs isotonic, 130-131
  isovolumic, 155, 159
  labor 992-993
  maximum strength of, 132
  micturition 461
  molecular mechanism of, 122-126
  of lungs 446-467
  power of, 1009, 1010
  premature, 175, 199-201
  segmentation, 765, 766
  series elastic component of, 130-131
  "servo" control of, 611
Contraction(s) (*Continued*)
  skeletal muscle, 131-133
  sliding mechanism of, 122-123
  smooth muscle, 140-141
    tonic contraction, 146, 756
    without action potentials, 145-146
  summation of, 131-132
  synchronous, Purkinje system in, 170
  tetanic, 336
  velocity of, 126
  "walk along" theory of, 124-125
Contrast, visual, 722, 726
Contrast enhancement, cortical control of, 591
  lateral inhibition in, 566-567, 588
Contrecoup, 374-375
Control systems, automaticity of, 7
  characteristics of, 5-6
  examples of, 5
  gain in, 6
  in premature infants, 1004
  negative feedback in, 6
  physical principles of, 7-10
  positive feedback in, 6
  symbols used in, 7-8
Conus medullaris, 803
Convection, heat loss by, 851
Convergence, of signals, 565
Convulsions, 570, 674
Cord See *Spinal cord*
Cord righting reflex, 616
Cordotomy, 596
Corona radiata, 971, 983
Coronary artery, blood flow control in, 296-299
  blood flow measurement in, 303-304
  in ischemic heart disease, 299-303
    current of injury from, 194-195
  normal blood flow in, 295-296
  spasm of, 299
Coronary circulation, control of, 296-299
  in ischemic heart disease, 299-303
  in shock, 327
  measurement of, 303-304
  normal, 295-296
Coronary shock, 278, 300
Coronary sinus, 295
"Coronary steal" syndrome, 302
Coronary thrombosis, recovery from, 195
Corpus albicans, 971
Corpus callosum, 655, 662-663
Corpus luteum, 971-972, 987-988
Correction, in control system, 6
Cortex, adrenal, 909-910
    hormones of See *Adrenocortical hormones*
    primary atrophy of, 920
  agranular, 651-652
  auditory, 653, 741-742
  cerebellar, 641-643
  cerebral See *Cerebral cortex*
  limbic, 676-677, 683
  motor See *Motor cortex*.
  of cytoplasm, 14
  olfactory, 751
  renal, 393
  somatic sensory, 584-587, 652-654
  temporal, 727
  visual, 653, 725-728, 730
Corticofugal pathways, 590-591
Corticomedial nuclei, 682
Corticopontocerebellar pathway, 640
Corticorubral tract, 636
Corticorubrospinal system, 636
Corticospinal tract, 632, 635, 636
Corticosteroids See *Adrenocortical hormones*.
Corticosterone, 910, 911, 914
Corticotropes, 885
Corticotropin. See *Adrenocorticotropic hormone (ACTH)*
Corticotropin-releasing factor (CRF), 886, 917-918, 919
Cortisol, 877, 914
  anti inflammatory effects of, 916-917
  chemical formula of, 910
  effect on ACTH secretion, 918
  effect on carbohydrate metabolism, 914-915
  effect on fat metabolism, 915-916
  effect on gluconeogenesis, 817, 914
  effect on protein metabolism, 915
  formation of, 910
  in stress, 916
  secretion of, 909-910
    excessive, 920-921
    loss of, 920
    regulation of, 917-919
  transport of, 911
Cortisol binding globulin, 911
Cortisone, 910, 914
Cough reflex, 475
Coumarins, 85
Countercurrent mechanism, 414-417
Coupled reactions, 808
Cramps, 598-599, 817
Creatinine, concentration of, 459
  in chronic renal failure, 454, 455
  reabsorption of, 405, 406
Cretinism, 907-908
CRF (corticotropin releasing factor), 886, 917-918, 919
Cricoarytenoid muscles, 478
Cricopharyngeal muscle, 760
Crisis, febrile, 859
Crista ampullaris, 622
Critical closing pressure, 214
Critical frequency, 132
Cross bridges, myosin, 121, 123, 124-125
Crossed extensor reflex, 614-615
Cross-eyedness, 731
Cryptorchidism, 958, 965
Crypts of Lieberkühn, 770, 784-785
CSF (cerebrospinal fluid), 374-377
  carbon dioxide concentration in, 507
CSF pressure, 376
  headache from low, 602-603
Cuboidal epithelioid cells, 897
Cumulus oophorus, 970
Cuneate nucleus, 583
Cuneocerebellar tract, 641
Cupula, 622
Curare, 138
Curariform drugs, 139
Current of injury, 192-195, 301
Cushing reaction, 251
Cushing's syndrome, 920-921
Cutaneous circulation, 343-345
Cyanide poisoning, 522
Cyanocobalamin (vitamin B_{12}), 869
  in fetus, 998
  in red blood cell formation, 45
Cyanosis, 323-324, 522
Cyclic 3',5'-adenosine monophosphate (cyclic AMP), ACTH effect on, 91
  formation of, 811
  in hormone mediation, 880-881
  in parathyroid stimulation, 916
  in platelets, 76
  in renal tubules, 416
  in smooth muscle contraction, 145

Cyclic 3',5'-adenosine monophosphate (cyclic AMP) (*Continued*)
in thyroid hormone secretion, 903
Cyclic guanosine monophosphate, as second messenger, 881
Cyclosporin A, 75
Cystic duct, 781
Cystinuria, essential, 458
Cystometrogram, 461
Cytochrome oxidase, 814
Cytochromes, 814
Cytoplasm, 11, 14-18
Cytosine, in DNA, 27, 28
in RNA, 29
Cytoskeleton, 16
Cytosol, 14
Cytotrophoblast cells, 985

Dale principle, 553
Damping, by cerebellum, 645
by stretch reflex, 610
Dark adaptation, 716-717
Dead space, 473
anatomic, 473
physiologic, 473, 491, 492, 518
in emphysema, 519-520
Deafness, 743-744
Deamination, 832-833, 837
Decarboxylases, 813
Deceleratory forces, 534
Decerebrate animal, 606, 620-621
Decibel, 740
Decidual cells, 985, 988
Decompensation, in heart failure, 308-309, 314-315
Decompression, of diver, 539-541
Decremental conduction, 558-559
Deep cerebellar nuclei, 641, 642
Deep sea diving, change in density of air in, 541
decompression in, 539-541
effect of rapid ascent in, 539-541 542
effect of rapid descent in, 541-542
effect on volume of gases, 537
high partial pressures in, 537-541
relation of depth to pressure in, 537
SCUBA, 542
sea level volume of air in, 541
Deep sensations, 580
Defecation, 768-769
control of, 802
paralysis of, 803
Defecation reflexes, 768-769
Defense pattern, 240-241
Defervescence, 859
Defibrillation, of ventricles, 174-175
Deflection plates, 118
Deglutition, 760-761
Dehydration, effect on extracellular fluid, 391
intracellular, 431
shock from, 332
7 Dehydrocholesterol, 938
Dehydroepiandrosterone, 920, 961, 988
Dehydrogenase, 812-813
Deiodinase enzyme, 899
Delta cells, 923, 932
Delta waves, 669, 670, 671-672
Dendrite(s), 550
role in synaptic transmission, 558-559
Denervation, muscle, 134
Denervation supersensitivity, 693-694
Dense bodies, 141
Dental lamina, 951
Dentate nucleus, 641
Dentinal tubules, 951
Dentine, 951
Dentition, 951-952
Deoxyadenylic acid, in DNA, 28
Deoxycytidylic acid, in DNA, 28
Deoxyguanylic acid, in DNA, 28
Deoxyribonuclease, 778-779
Deoxyribonucleic acid (DNA), 12
in mitochondrion, 17-18
in nucleus, 18
proofreading of, 36
repair of, 36
replication of, 35-36
structure of, 27-28
transcription of 29-33
effect of growth hormone on, 887-888
effect of insulin on 928
effect of thyroid hormones on, 900
Deoxyribose, in DNA, 27, 28
Deoxythymidylic acid, in DNA, 28
Depolarization, atrial, 187
in gastrointestinal smooth muscle, 756
membrane, 106
ventricular, 185-187
without action potentials, 143
Depolarization waves, 177-178
prolonged, 196
slow conduction of, 195-196
Depression, norepinephrine and serotonin systems in 683-684
Depth of focus, 705
Depth perception, 708 725
Dermatomes, 591
Desmolase, 918
Desoxycorticosterone, 910, 911
Detrusor muscle, 460, 461
Deuteranope, 718-720
Development, behavioral, 1005
fetal, 997-999
of child, 1005
of premature infant, 1004
Dexamethasone, 910
Dextran solution, in shock, 334
α Dextrinase, 788
Dextrins, α limit, 787
Diabetes, adrenal, 914-915
pituitary 889
Diabetes insipidus, 427, 694
nephrogenic, 458
Diabetes mellitus, atherosclerosis with, 827
coma in, 928, 935
diagnosis of, 934-935
etology of, 933
fat metabolism in, 823, 825
from growth hormone, 888
in giantism, 891
infants of mothers with, 1004
juvenile, 933
ketosis in, 822
maturity-onset, 933
pathological physiology of, 933-934
respiratory acidosis from, 449
treatment of, 935
vs pituitary diabetes, 889
Dialysis, 455-457
Diamox, 449, 460
Diapedesis, 53, 56
Diaphragm, 466
Diarrhea, 765, 802-803
metabolic acidosis from, 448
Diastasis, 155
Diastole, 153
in clinical usage, 157
ventricles in, 155, 295
Dielectric, 104
Diet, for endurance training, 1009
high fat, 822, 827
of diabetic, 935
saturated fat, 825, 827
taste preference in control of, 748
unsaturated fat, 825
Dietary balances, 861-863
Differentiation, of cells, 33, 37-38
Diffuse junctions, 144
Diffusing capacity, at high altitudes, 531
diseases that affect, 518, 519
of athletes, 1014-1015
of respiratory membrane, 489-490
Diffusion, carrier-mediated, 92-93
cellular, 19
defined, 89
facilitated, 92-93
in renal tubules, 401
of glucose, 810, 925-926
gaseous. See *Gaseous diffusion*.
gastrointestinal absorption by, 792, 794
membrane potentials caused by, 101-103
net rate of, 93-94, 481, 483
of oxygen. See *Oxygen diffusion*.
of water, net, through cell membrane, 94-97
through capillary membrane, 350
through lipid bilayer, 90
simple, 89-90
through capillary membrane, 3, 350-351
through cell membrane, 89-93
through placental membrane, 986-987
vs active transport, 89
vs filtration, 358
Diffusion coefficient, 483, 488
Diffusion gradient, 483
Diffusion potential, 101-103, 105
Digestion, 787-790
by cell, 20-21
in newborns, 1002-1003
Digestive vesicle, 20, 54
Digitalis, effect on T wave, 196
for decompensated heart disease, 309, 314
Diglycerides, 795
Dihydrotestosterone, 960, 961, 963
1,25 Dihydroxycholecalciferol, 871, 938-939
Diiodotyrosine, 898, 899
Diisopropyl fluorophosphate, 139, 696
Diluting segment, of distal tubule, 402, 403
Dilution principle, 384-385
Diodrast, 408
Diopter, 702-703
Dipalmitoyl lecithin, in alveolar surfactant, 467
Dipeptidases, 790
2,3-Diphosphoglycerate (DPG), 498
Dipole layer, 104
Disaccharides, 787, 788, 810
Discharge zone, 564
Disseminated intravascular clotting, 84, 333
Dissociation, degree of, 440
Dissociation curves, carbon dioxide, 501
oxygen hemoglobin, 496, 498-499
Distal tubule, 394, 400
diluting segment of, 402, 403
late, 403, 417, 419, 420
osmotal concentration in, 417
osmosis through, 401
potassium excretion by, 420, 421

Distal tubule (*Continued*)
potassium reabsorption by, 420, 421
sodium reabsorption by, 419
sodium transport through, 419-420
Distension, and peristalsis, 758
vascular, 214-215
Distractibility, 656
Diuresis, osmotic, 934
water, 262, 430-431
Diuretics, 459-460
alkalosis caused by, 449
for essential hypertension, 270
for cardiac failure, 309
loop, 459-460
osmotic, 459
Divergence, signal, 564-565
Diver's paralysis, 540
Diving. See *Deep sea diving*
DNA. See Deoxyribonucleic acid (DNA)
DNA ligase, 36
DNA polymerase, 36
Dominant hemisphere, 654-655
Donnan effect, 356-357
L-Dopa 630
Dopamine, 553
in neostriatum-substantia nigra pathway, 627, 630
in schizophrenia, 684
Dorsal column-lemniscal system, 582
anatomy of, 583-584
in position sense, 588-589
sensations transmitted by, 582-583
somatic association areas in, 586-587
somatic sensory cortex in, 584-586
transmission in, 587-588
Dorsal horns, pain inhibitory complex in, 597
Dorsal motor nucleus, of vagus nerve, 239
Dorsal respiratory group, 504-505
Dorsal spinocerebellar tract, 640-641
Down-regulation of receptors, 879
dP/dt record, 163-164
DPG (2,3-diphosphoglycerate), 498
Drinking, temporary relief of thirst by, 432
threshold for, 432
Drugs, acetylcholinergic, 630
adrenergic 696
antidepressant, 684
antihistaminic, 335
cardiotonic 309
cholinergic blockers, 696
curariform, 139
for myasthenia gravis, 139
ganglionic blocking, 696-697
muscarinic, 696
neuromuscular junction and, 138-139
nicotinic, 696-697
parasympathomimetic, 696
sympathetic blocking, 303
sympatholytic, 334
sympathomimetic, 334, 335, 690, 696
thrombocytopenia from sensitivity to, 84
tranquilizer, 681
used by athletes, 1018
vasoconstrictor, 241-242
vasodilator, 242-243, 270, 302-303
Ductus arteriosus, 322, 1000 See also *Patent ductus arteriosus*
Ductus venosus, 1000, 1001
Duodenocolic reflex, 768
Duodenum, control of stomach emptying by 764-765
in intestinal phase of gastric secretion, 778
Duodenum (*Continued*)
ionic absorption in, 793
Dwarfism, 891
Dynein, 26
Dynorphin, 597
Dysarthria, 647
Dysbarism, 540
Dysdiadochokinesia, 647
Dyslexia, 654, 662, 728
Dysmenorrhea, 601
Dysmetria, 647
Dyspnea, 522-523
emotional, 523
in asthma, 521
in low cardiac reserve, 312-313
neurogenic, 523

Ear, 734-740
Eardrum, 734-735
Eating, 759-762
Eclampsia, 991
Ectopic focus, 175, 199-200
Ectoplasm, 14, 16
in ameboid locomotion, 24-25
Eddy currents, 210
Edema, brain, 340, 514
brawny, 368
cardiac, 311-312, 368
cerebral, 532
defined, 366
extracellular, 455
free fluid in, 352
from abnormal capillary dynamics, 368-370
from kidney retention of fluids, 370
gel-matrix in, 370-371
hypoproteinemic, 1004
in cardiac failure, 308, 309, 311-312, 368
in chronic renal failure, 454-455
in nephrotic syndrome, 458
interstitial fluid pressure in, 366-368
intra-alveolar, 292
intracellular, 455
nonpitting, 368, 916
nutritional, 369
peripheral, 311, 368
pitting, 368
pulmonary. See *Pulmonary edema*
safety factors in, 292-293, 368, 372
stretch of tissue spaces in chronic, 367-368
Edinger-Westphal nucleus, 732, 733
Effectors, 547
receptors of, 689-690
Efferent arterioles, 393, 394
blood flow through, 395
blood pressure in, 396
in glomerular filtration rate, 399-400, 412-413
Efficiency, of heart, 157
Einthoven's law, 181
Einthoven's triangle, 181
Ejaculation, 959-960
Ejaculatory duct, 954
Ejection fraction, 155
Ejection period, 155, 159
Ejection pressure, 157
Elastic fibers, in lungs, 467
Elastic modulus, 317
Elastic recoil, 466
Elastic work, 469
Electrical alternans, 199
Electrical axis, of heart, 188-191
Electrical neutrality, principle of, 104
Electrical potential(s), in cardiac cycle, 179-180
inside neuronal soma, 555
vectorcas, 183
Electrical potential difference, across neuronal somal membrane, 554
effect on diffusion, 94
Electrocardiogram(s), decreased voltage of, 191
in cardiac arrhythmias, 197-203
in cardiac cycle, 154-155, 179-180
leads in, 180-182
methods for recording, 179
normal, 177-179
time calibration of, 178-179
vectorial analysis of, 183-188
voltage calibration of, 178-179
zero reference potential of, 193
Electrochemical gradient, in renal tubule, 401
Electroconvulsive therapy, 684
Electrode, indifferent, 103, 181
Electroencephalogram, 669-670
in epilepsy, 674
in sleep and wakefulness, 671
Electrogenic pump, 103, 104-106
Electrolytes, estrogen effect on, 975
in cell, 11
progesterone effect on, 975
secretion in gastrointestinal tract, 771-772, 796
testosterone effect on, 962-963
Electromotive force, 94
Electromyogram, 134-135
Electron transport chain, 814
Electro-olfactogram, 749
Electrotonic current, 558-559
Elephantiasis, 369
Embolism, air, 293, 542, 543
fat, 293
pulmonary, 84, 293
Embolus(s), 84, 293, 299
Embryo, See also *Fetal, Fetus.*
implantation of, 984-985
intrauterine nutrition of, 983
primary organizer of, 38
Emission, 959-960
Emmetropia, 705
Emphysema, blood flow in, 293-294
chronic pulmonary, 518-520
electrocardiogram in, 191
respiratory membrane in, 488
ventilation perfusion ratio in, 491-492
Enamel, 951
End-diastolic pressure, 158
End-diastolic volume, 155, 159
End feet. See *Terminal(s).*
Endocochlear potential, 739
Endocrine glands, in infant, 1003-1004
overview of, 876-878
Endocytosis, 19-20
Endolymph, 622, 739
Endometrial cycle, 975-979
Endometriosis, 982
Endoplasm, 14
in ameboid locomotion, 24-25
Endoplasmic matrix, 15
in protein synthesis, 32
Endoplasmic reticulum, 14-15
agranular, 15, 21
granular, 15, 21
ribosome attachment to, 15, 32
role in cell synthesis, 21-22
smooth, 15, 21
vesicles of, 21-22

Endorphins, 886
in pain control system, 597, 684
secretion with ACTH, 919
Endotoxins, disseminated intravascular clotting from, 84
in fever, 858
in shock, 330, 333
released by ischemic tissues, 329-330
End plate, 136
End plate potential, 137-138
End systolic volume, 155, 159
Endurance, 1009, 1010
Energy, aerobic, 842-843
anaerobic, 815-816, 842-843
ATP in, 841-842
cellular, 22-24, 843-844
coupled reactions for, 808
for cardiac contraction, 157
free, 808
in foods, 808, 861-863
in obesity, 865
phosphocreatine as storage depot for, 842
proteins in, 832-833
release of, cellular control of, 843-844
from glucose, 811-818
triglyceride control of, 823-824
requirements for various activities, 845-848
triglycerides in, 820-823
Enkephalins, 886
as neurotransmitters, 553
in pain control system, 597, 684
Enteric lipase, 789
Enteric nervous system, 756-758, 770
Enteritis, 785, 802
Enterogastric reflex, 784, 778
Enterogastrone, 765
Enterohepatic circulation, 783
Enterokinase, 779
Enzyme(s), 11, 12, 27
cascade of, 880, 931
converting, 242, 254
effect of thyroid hormones on, 900
gastric, 775
hydrolytic, 16-16
in metabolic reactions, 843-844
in small intestine, 785
oxidative, 23
pancreatic, 779
proteolytic, in blood coagulation, 80
in phagocytosis, 54
in saliva, 773
in sperm, 955, 958, 983
Enzyme activation, 34-35
Enzyme inhibition, 34, 35
Enzyme regulation, 33, 34-35
Eosinophil(s), effect of cortisol on, 917
in infections, 51, 57-58
Eosinophil chemotactic factor, 58, 521
Ependyma, 377
Ephedrine, 696
Epididymis, 954
in spermatogenesis, 956-957
Epilepsy, 674-675
neuronal circuits in, 570
Epinephrine, as adrenergic drug, 696
derivation of, 878
in control of bronchioles, 475
in cutaneous blood flow regulation, 344-345
in fat utilization, 824, 930
in hypertension, 267
in muscle blood flow regulation, 337
in vasoconstrictor system, 241, 253
phosphorylase activation by, 811
Epinephrine (*Continued*)
release of, 692-693, 876
Epithalamus, 676
EPSP (excitatory postsynaptic potential), 555
summation of, 558
time course of, 557
Equatorial plate, 37
Equilibrium, cerebellum in, 643-644
in blind flying, 534
maintenance of, 621-625
Equilibrium point, in arterial pressure control, 258-260
in cardiac output and venous return, 282
Erection, female, 980
male, 959
Ergocarciferol, 871
Erythema, 916
Erythremia, 49
Erythroblast(s), 43, 45
Erythroblastosis fetalis, 49, 72-73
jaundice in, 1002
spleen in, 343
Erythrocytes See *Red blood cells.*
Erythrogenesis, 42-45
Erythropoietin, 44-45, 455
Esophageal sphincter, lower, 761-762
upper, 780
Esophageal varicosities, 342
Esophagus, disorders of, 798
in swallowing, 761
pain in, 600
secretion in, 774
Estradiol, 956, 972-977
Estriol, 972
Estrogen(s), 877, 972
and athletic performance, 1008, 1009
and blood cholesterol, 825
chemistry of, 972
degradation of, 972
effect on *FSH* and *LH*, 977-978
effect on sexual characteristics, 974-975
effect on uterine contractility, 991-992
in breast development, 975, 994
in endometrial cycle, 976
in libido, 684
in menopause, 979
in spermatogenesis, 956
intracellular functions of, 975
produced by ovaries, 968
produced by testes, 961
secreted by corpus luteum, 972
secreted by placenta, 988
suppression of fertility by, 981
synthesis of, 972
transport of, 972
Estrone, 972
Ethacrynic acid, 459-460
Ethynodiol, 981
Ethynyl estradiol, 981
Eunuchism, female, 979
male, 965
Eupnea, 516
Evans blue dye, 385
Evaporation, heat loss by, 851
Excitability curve, of nerve fibers, 116-117
Excitation, electrical events during, 553-556
inhibition of, 117, 552
mechanisms of, 115-117
of action potential, 115-117
of cardiac muscle, 165-171
of skeletal muscle fibers, 127, 137-138
of smooth muscle, 143, 144-146
Excitation (*Continued*)
receptors in, 550, 552
refractory period for, 117
role of dendrites in, 558-559
threshold for, 116, 556, 559-560
Excitation-contraction coupling, in cardiac muscle, 152-153
in cardiac failure, 308-309
in skeletal muscle, 126-129
in smooth muscle, 143-144
Excitatory postsynaptic potential (EPSP), 555
summation of, 558
time course of, 557
Excitatory receptors, 550, 552
in smooth muscle, 144, 145
Excitatory state, 559-560
Excitatory stimulus, 564
Excited zone, 564
Excitement, hypothalamus in, 681
Excretion, 4
fluid volume, 421-424
of calcium, 937, 945
of excess solutes, 414-417
of excess water, 413-414
of phosphate, 937, 945
of potassium, 420-421
of sodium, 419-420
of urea, 417-419
through placental membrane, 987
Exercise, aerobic energy during, 842-843
arterial pressure during, 338
blood flow during, 336, 337-338
body fluids and salt in, 1017-1018
body heat in, 1017
bradycardia and, 197
cardiac output regulation in, 272, 273, 276-277, 283-284, 338
cardiovascular system in, 1015-1017
dissolved oxygen in, 499-500
effect on energy output, 865
effect on glucagon secretion, 932
effect on metabolic rate, 845-848
mean circulatory filling pressure during, 252
muscle vasodilatation in, 276
muscles in, 1009-1014
obesity and, 865, 866
oxygen transport during, 497
oxygen uptake during, 494
oxygen-diffusing capacity during, 489
oxygen-hemoglobin dissociation curve during, 498-499
pulmonary circulation in, 290-291
respiration in, 1014-1015
respiratory control during, 511-513
sympathetic nervous system in, 276-277, 337-338
water loss during, 383
with diabetes, 935
with valvular lesions, 321
Exercise test, 312-313
Exocytosis, by cells, 20, 22
of chylomicrons, 795
of secretory vesicles, 771
of synaptic vesicles, 551
Exophthalmos, 905-906
Expansion, of lungs, 466-467
Expiration, mechanics of, 466-467
ventral respiratory group in, 505-506
work of, 469
Expiratory reserve volume, 470, 471
Extensor spasm, 638
Exteroceptive sensations, 580, 625
Extracellular fluid, calcium in, 937-940, 944-946

Endorphins, 886
in pain control system, 597, 684
secretion with ACTH, 919
Endotoxins, disseminated intravascular clotting from, 84
in fever, 858
in shock, 330, 333
released by ischemic tissues, 329–330
End plate, 136
End plate potential, 137–138
End-systolic volume, 155, 159
Endurance, 1009, 1010
Energy, aerobic, 842–843
anaerobic, 815–816, 842–843
ATP in, 841–842
cellular, 22–24, 843–844
coupled reactions for, 808
for cardiac contraction, 157
free, 808
in foods, 808, 861–863
in obesity, 865
phosphocreatine as storage depot for, 842
proteins in, 832–833
release of, cellular control of, 843–844
from glucose, 811–816
triglyceride control of, 823–824
requirements for various activities, 845–846
triglycerides in, 820–823
Enkephalins, 886
as neurotransmitters, 553
in pain control system, 597, 684
Enteric lipase, 789
Enteric nervous system, 756–758, 770
Enteritis, 785, 802
Enterogastric reflex, 784, 778
Enterogastrone, 765
Enterohepatic circulation, 783
Enterokinase, 779
Enzyme(s), 11, 12, 27
cascade of, 880, 931
converting, 242, 254
effect of thyroid hormones on, 900
gastric, 775
hydrolytic, 15–16
in metabolic reactions, 843–844
in small intestine, 785
oxidative, 23
pancreatic, 779
proteolytic, in blood coagulation, 80
in phagocytosis, 54
in saliva, 773
in sperm, 955, 958, 983
enzyme activation, 34–35
enzyme inhibition, 34, 35
enzyme regulation, 33, 34–35
eosinophil(s), effect of cortisol on, 917
in infections, 51, 57–58
eosinophil chemotactic factor, 58, 521
ependyma, 377
ephedrine, 696
epididymis, 954
in spermatogenesis, 956–957
epilepsy, 674–675
neuronal circuits in, 570
epinephrine, as adrenergic drug, 696
derivation of, 878
in control of bronchioles, 475
in cutaneous blood flow regulation, 344–345
in fat utilization, 824, 930
in hypertension, 267
in muscle blood flow regulation, 337
in vasoconstrictor system, 241, 253
phosphorylase activation by, 811

Epinephrine (*Continued*)
release of, 692–693, 876
Epithalamus, 676
EPSP (excitatory postsynaptic potential), 555
summation of, 558
time course of, 557
Equatorial plate, 37
Equilibrium, cerebellum in, 643–644
in blind flying, 534
maintenance of, 621–625
Equilibrium point, in arterial pressure control, 258–260
in cardiac output and venous return, 282
Erection, female, 980
male, 959
Ergocarciferol 871
Erythema, 916
Erythremia 49
Erythroblast(s), 43, 45
Erythroblastosis fetalis, 49, 72–73
jaundice in, 1002
spleen in, 343
Erythrocytes See *Red blood cells*
Erythrogenesis, 42–45
Erythropoietin, 44–45, 455
Esophageal sphincter, lower, 761–762
upper, 760
Esophageal varicosities, 342
Esophagus, disorders of, 798
in swallowing, 761
pain in, 600
secretion in, 774
Estradiol, 956, 972–977
Estriol, 972
Estrogen(s), 877, 972
and athletic performance, 1008, 1009
and blood cholesterol, 825
chemistry of, 972
degradation of, 972
effect on FSH and LH, 977–978
effect on sexual characteristics, 974–975
effect on uterine contractility, 991–992
in breast development, 975, 994
in endometrial cycle, 976
in libido, 684
in menopause, 979
in spermatogenesis, 956
intracellular functions of, 975
produced by ovaries, 968
produced by testes, 961
secreted by corpus luteum, 972
secreted by placenta, 988
suppression of fertility by, 981
synthesis of, 972
transport of, 972
Estrone, 972
Ethacrynic acid, 459–460
Ethynodiol, 981
Ethynyl estradiol, 981
Eunuchism, female, 979
male, 965
Eupnea, 516
Evans blue dye, 385
Evaporation, heat loss by, 851
Excitability curve, of nerve fibers, 116–117
Excitation, electrical events during, 553–556
inhibition of, 117, 552
mechanisms of, 115–117
of action potential, 115–117
of cardiac muscle, 163–171
of skeletal muscle fibers, 127, 137–138
of smooth muscle, 143, 144–146

Excitation (*Continued*)
receptors in, 550, 552
refractory period for, 117
role of dendrites in, 558–559
threshold for, 116, 556, 559–560
Excitation-contraction coupling, in cardiac muscle, 152–153
in cardiac failure, 308–309
in skeletal muscle, 126–129
in smooth muscle, 143–144
Excitatory postsynaptic potential (EPSP), 555
summation of, 558
time course of, 557
Excitatory receptors, 550, 552
in smooth muscle, 144, 145
Excitatory state, 559–560
Excitatory stimulus, 564
Excited zone, 564
Excitement, hypothalamus in, 681
Excretion, 4
fluid volume, 421–424
of calcium, 937, 945
of excess solutes, 414–417
of excess water, 413–414
of phosphate, 937, 945
of potassium, 420–421
of sodium, 419–420
of urea, 417–419
through placental membrane, 987
Exercise, aerobic energy during, 842–843
arterial pressure during, 338
blood flow during, 336, 337–338
body fluids and salt in, 1017–1018
body heat in, 1017
bradycardia and, 197
cardiac output regulation in, 272, 273, 278–277, 283–284, 338
cardiovascular system in, 1015–1017
dissolved oxygen in, 499–500
effect on energy output, 865
effect on glucagon secretion, 932
effect on metabolic rate, 845–846
mean circulatory filling pressure during, 252
muscle vasodilatation in, 276
muscles in, 1009–1014
obesity and, 865, 866
oxygen transport during, 497
oxygen uptake during, 494
oxygen-diffusing capacity during, 489
oxygen hemoglobin dissociation curve during, 498–499
pulmonary circulation in, 290–291
respiration in, 1014–1015
respiratory control during, 511–513
sympathetic nervous system in, 276–277, 337–338
water loss during, 383
with diabetes, 935
with valvular lesions, 321
Exercise test, 312–313
Exocytosis, by cells, 20, 22
of cholinergic, 795
of secretory vesicles, 771
of synaptic vesicles, 551
Exophthalmos, 905–906
Expansion, of lungs, 466–467
Expiration, mechanics of, 466–467
ventral respiratory group in, 505–506
work of, 468
Expiratory reserve volume, 470, 471
Extensor spasm, 638
Exteroceptive sensations, 560, 625
Extracellular fluid, calcium in, 937–940, 944–946

Extracellular fluid (*Continued*)
carbon dioxide concentration in, 5
components of, 3, 88, 386-387
defined, 2-3, 88
effect of aldosterone on, 911-912
effect of saline solution on, 391
effect of water on, 390-391
in arterial pressure regulation, 259-262
in blood volume control, 426
in hypertension, 265
measurement of, 386
nutrients in 4, 88
osmolality of, 428-433
oxygen concentration in, 5
phosphate in, 910, 944-946
potassium concentration in, 433-435
shifts between intracellular fluid and, 387-392
sodium concentration in, 428-433, 894
transport of, 3-4
volume of, control of, 428
vs intracellular fluid, 3, 88
Extracellular fluid compartment, 383
Extracorporeal circulation, 324
Extraction ratio, 408
Extrafusal muscle fibers, 608
Extrapyramidal system, 636
Extrasystole, 175
electrocardiogram in, 199-201
Eye(s), as camera, 703-704
autonomic control of, 690
focusing of, 732
headache from disorders of 603
movements of, control of, 634, 728-729
fixation, 729-731
optics of, 703-708
reduced, 703-704
repressed, 731
stabilization of, 623, 625

Face recognition, 655-656
Facial nerve, 747
Facilitated diffusion, 92-93
in renal tubules, 401
of glucose, 810, 925-926
Facilitated zone, 564
Facilitation, of neurons, 558, 564
centrifugal control of, 566
in memory formation, 548, 658-660
post tetanic 560, 658
presynaptic, 658
F-actin, 124
Factors, blood-clotting, 79, 80-81, 83
vs hormones, 886
Fahraeus Lindqvist effect 207
Fainting, emotional, 332
in cardiac failure, 305, 306
neurogenic shock from, 332
Fallopian tubes, 968
effect of estrogens on, 974
entry of ovum into, 983
progesterone effect on, 975
transport of ovum in, 981
Far-sightedness, 705-706
Fasciculation, muscle, 135
Fast channels, in action potential, 110, 113
in cardiac muscle, 151
in sinoatrial node rhythmicity, 166
Fastigial nucleus, 640, 641
Fasting blood glucose level, 931
Fat(s) See also *Lipid(s)*.
absorption of, 795-796

Fat(s) (*Continued*)
as protein sparers, 862
brown, 846, 855
carbohydrate formation from, 816-817
composition of, 787
depletion of, 866
digestion of, 788, 790
homogenized, 391-392
hormonal regulation of utilization of, 824
in different foods, 861
metabolism of See *Lipid metabolism*
neutral See *Triglycerides*
saturated, 825
stomach emptying and, 764-765
storage of, abnormalities in, 865
effect of insulin on, 926-927
importance of, 823
synthesis of, from glucose, 816
hydrogen in, 816
importance of, 823
unsaturated, 825
Fat cells, 820
Fat deposits, 820
effect of estrogens on, 974-975
Fat depots, 820
Fat embolism, 293
Fat sparer, 926
Fatigue, at neuromuscular junction, 138
muscle, 133, 1011
of reflexes 614, 615
synaptic 560
in duration of reverberation, 568
in nervous system stabilization, 670-571
Fatty acids, absorption of 795, 796
conversion of acetyl CoA into, 822
degradation of, 820-821
diffusion into fetal blood, 987
entry into mitochondria, 820
free, 789, 795, 819
metabolism of, 22
mobilization by cortisol, 915-916
nonesterified, 819
plasma, 928
structure of, 787, 818
transport of, 819
triglyceride formation from, 822-823
Fear, 680, 684
Feces, formation of, 796
Feedback, aldosterone, 427-428
gastrointestinal, 864
in control system, 8, 34
in ileocecal sphincter control, 767
in iron absorption, 47
in spermatogenesis, 964
in testosterone secretion 964
negative, in control systems, 6
in glomerular filtration rate, 412-413
in hormone secretion, 879
of operon, 31
osmosodium receptor-antidiuretic hormone system, 429-431
positive, in brain edema, 340
in control system, 6
in progressive shock, 328, 330
sensory, of motor functions, 637, 648-649
tubuloglomerular, 411-413
Feedback gain, for body temperature control, 856
for pressure control, 257
in control system, 6
in respiratory center, 615
in sensory pathways, 591

Feeding center, 863
Femoral thrombosis, 84
Fenestrae, 349, 397
Fenn effect, 125
Ferritin, 47, 838
Fertility, female, 981-982
male, 957-958
seasonal, 966
Fertilization, 958, 983-984
Fetal zone, 988
Fetus, blood supply to, 985-987
growth and development of, 997-999
human chorionic gonadotropin in, 963, 969, 988
testosterone in, 961
Fever, 858
and metabolic rate, 846
characteristics of, 858-859
hypothalamic thermostat in, 858
rheumatic, 67, 318
Fibrillation, 172
atrial, 174
electrocardiogram of, 202-203
in mitral valvular disease, 321
chain reaction mechanism of, 173-174
muscle, 135
Purkinje system and, 174
ventricular, 174-175
after myocardial infarction, 302
circulatory arrest from, 335
electrocardiogram in, 203
vulnerable period for, 173-174
Fibrin, antithrombin action of, 82
formation of, 78-79
Fibrinogen, conversion to fibrin, 78-79
in seminal vesicles, 957
plasma, 208, 831
Fibrinolysin, 82, 83, 976
Fibrin-stabilizing factor, 76, 78
Fibroblasts, in blood clot formation, 77
Fick procedure, 284
Fields of Forel, 211
Fields of vision, 729
Fifth nerve, 747
Fight or flight reaction, 695
Filariasis 369
Filtration, glomerular, dynamics of, 398-399
in cardiac failure, 306
net 359
vs diffusion, 358
Filtration coefficient, of capillary exchange, 359
renal, 399
Filtration fraction, from plasma clearances, 408
of glomerular filtration, 398
Filtration pressure, at capillary membrane, 358
pulmonary, 371-372
renal, 399
Fixation movements, 729-731
Flagellum, 25
Flatus, 796, 805
Flavin adenine dinucleotide, 868
Flavin mononucleotide, 868
Flavoprotein, 814
Flexing movements, 626
Flexor reflex, 613-614
Flocculonodular lobes, 623, 639, 640
Flower spray ending, 608
Flowmeters, 208, 284
Fluid balance, neonatal, 1002
Fluid exchange See *Capillary exchange*
Fluid retention, 262-263
edema from, 370

Fluid retention (*Continued*)
effect of testosterone on, 962-963
in cardiac failure, 306-307, 308, 309, 311-312
in chronic renal failure, 454-455
in Goldblatt hypertension, 268-269
Fluid space, 386
Fluid volume, changes in, 387-392
in arterial pressure regulation, 259-262
in hypertension, 265
measurement of, 384-386
regulation of, 421-424
Fluids, body. See *Body fluids*
Fluorine, 872, 953
9α-Fluorocortisol, 910
Fluorosis, 872
Flush, in fever, 859
Flutter, 172
atrial, 172-173, 202
Focal length, 702
Focal line, 701
Focal point, 701
Folds of Kerckring, 791
Folic acid, 869
in fetus, 998
in red blood cell maturation, 45
in white blood cell maturation, 52
Folium, 641
Follicle(s), ovarian, 968-970, 978
thyroid, 897
Follicle-stimulating hormone (FSH), 877, 884
in follicular growth, 968-970, 977, 978
in male sexual function, 963-964
in menopause, 979
in spermatogenesis, 956
Follicle-stimulating hormone releasing hormone (FSHRH), 977
Food(s), digestion of, 762-765, 787-790
energy available in, 861-863
energy release from, 808
ingestion of, 759-762
regulation of intake of, 863-866
Food stores, depletion of, 866
Foramen, of Luschka, 375
of Magendie, 375
of Monro, 375
ovale, 1000, 1001
Forced expiratory vital capacity, 477
Forced expiratory volume, 477
Forward progression, 626
Fovea, 708, 711, 720
Frank-Starling law, 157-160, 273, 338
Free fatty acid, 789, 795, 819
Free fluid, in interstitium, 352
Free fluid pressure, 354-355
in edema, 371
in lungs, 371
in lymph flow, 363
positive, 367-368
Free fluid spaces, 370
Free fluid vesicles, 352
Free nerve endings, 562, 580
as pain receptors, 593
Free water clearance, 417
Freon toxicity, in submarine, 543
Frequency, of sound, 737-738, 739, 742
Frohlich's syndrome, 965
Frostbite, 860
Fructose, absorption of, 788, 794
in seminal vesicles, 957
FSH See *Follicle stimulating hormone (FSH)*.
FSHRH (follicle-stimulating hormone releasing hormone), 977
Functional residual capacity, 471, 472, 484
in asthma, 521
in newborn, 1001
Furosemide, 459-460
Fusiform layer, 651
Fusion, of flickering light, 717
of visual images, 725, 731

G cells, 776
G force, 532-533
GABA (gamma-aminobutyric acid), 553
in neostriatum-substantia nigra pathway, 627, 629
G-actin, 124
Gain See *Feedback gain*
Galactose, 788, 794
Gallbladder, 781-782
pain in, 600
Galloping reflex, 616
Gallstones, 783-784, 801
Gamma chains, 46
Gamma globulins, 208
Gamma motor nerve fibers, 608
brain areas for control of, 612
control of static and dynamic responses by, 609
in voluntary motor activity, 611
Gamma motor neurons, 607
Gamma aminobutyric acid (GABA), 553
in neostriatum-substantia nigra pathway, 627, 629
Ganglion cells, 720-725
Ganglionic blocking drugs, 696-697
Gangrene, gas, 543
shock from, 333
Gap junctions, 140, 549
in gastrointestinal wall, 755
Gas concentration, in alveoli, 485, 490-492
in fluid, 482
Gas gangrene, 543
Gas pressures, 482-483, 490
Gas volumes, actual vs sea level, 537, 541
effect of depth on, 537
Gaseous diffusion, capacity for, 489-490
concentration gradient in, 481
molecular basis of, 481
net, 483
physics of, 481-483
pressure gradient for, 483
through fluids, 483
through respiratory membrane, 482, 488-490
through tissues, 483
ventilation perfusion ratio and, 490-492
Gases, blood, abnormalities of, 514-515
during exercise, 511-513, 1015
in respiratory control, 507-511
measurement of, 517
body, effect of high partial pressures on, 537-541
Gastric. See also *Stomach*.
Gastric amylase, 775
Gastric atony, 801
Gastric atrophy, 799
Gastric barrier, 799
Gastric glands, 762, 774
Gastric inhibitory peptide, 758, 765, 778
Gastric lipase, 775, 788
Gastric mucosa, in pernicious anemia, 45
Gastric pain, 600
Gastric secretion, 774-778
Gastrin, chemical composition of, 777
effect on ileocecal sphincter, 767
effect on stomach emptying, 763-764
in gastrointestinal blood flow control, 342
regulation of pancreatic secretion by, 780-781
secretion of, 774, 776
Gastritis, 798-799
Gastrocolic reflex, 768
Gastroenteric reflex, 766
Gastroesophageal sphincter, 761-762
Gastroileal reflex, 766
Gastrointestinal absorption, anatomical basis of, 790-792
effect of aldosterone on, 912
in large intestine, 795
in premature infant, 1004
in small intestine, 792-796
mechanisms of, 792
neonatal, 1002-1003
of calcium, 793, 937-939, 945-946
of phosphate, 937, 939, 945-946
Gastrointestinal glands, 770-771
Gastrointestinal hormones, 771
effect on insulin secretion, 930
Gastrointestinal reflexes, 757-758
defecation, 768-769
duodenocolic, 768
enterogastric, 764
gastrocolic, 768
gastroenteric, 766
gastroileal, 766
peristaltic, 769
peritoneointestinal, 769
renointestinal, 769
somatointestinal, 769
vesicointestinal, 769
Gastrointestinal secretion, by esophagus, 774
by large intestine, 795
by pancreas, 778-781
by small intestine, 784-785
by stomach, 774-778
of bile, 781-784
of electrolytes, 771-772, 796
of saliva, 772-774
of water, 771-772, 779-780
Gastrointestinal tract, 754
amino acids absorbed from, 829-830
autonomic control of, 691, 694, 757-758
blood flow through, 342
digestion in, 787-790
disorders of, 798-805
effect of thyroid hormone on, 902
feedback to feeding center, 864
fetal, 998
gases in, 805
hormonal control of, 758
in ingestion of food, 759-762
innervation of, 756-758
iron absorption from, 47
lipid transport from, 818-819
motility of, 754-758
mucus in, 772
nutrients supplied by, 4
obstruction of, 804-805
shock from, 331
thiamine deficiency and, 868
types of movements in, 758-759
wall of, 754-756
Gating, of protein channels, 91-92, 137
Gel fluid, 352
Gelatinase, 775

Growth hormone (GH) (*Continued*)
metabolic effects of, 887–889
physiological functions of, 887–893
secretion of, abnormalities of, 891–893
regulation of, 889–891
Growth hormone inhibitory hormone (GHIH). See *Somatostatin.*
Growth hormone releasing hormone (GHRH), 886, 890
Guanethidine, 696
Guanine, in *DNA*, 27, 28
in RNA, 29
Gut. See *Large intestine; Small intestine*
Gynecomastia, 966

H zone, 122
Habituation, 659, 681
Hair, end-organ of, 581
estrogen effect on, 975
testosterone effect on, 961–962
Hair cells, in auditory system, 735, 738–739
in vestibular apparatus, 621–622
Haldane effect, 501–502
Hand movements, control of, 634
Haptens, 61
Haptoglobin, 74
Haustration, 767
Haversian canal, 943
Hay fever, 68
Head factors, for feeding, 864–865
Head orientation, detection of, 621–622
Head rotation, control of, 634
detection of, 623–625
Headache, 602–603
Head-down position, 334
Hearing, abnormalities of, 743–744
central auditory mechanisms, 740–743
cochlea in, 735–740
frequency range of, 740
ossicular system in, 734–735
threshold for, 740
tympanic membrane in, 734–735
Heart. See also *Cardiac.*
abnormal rhythms of, 171–176
collateral circulation in, 299–300
conduction in, 163–171
congenital anomalies of, 321–324
effectiveness of, 279
excitation of, 165–171
Frank-Starling law of, 157–160, 273, 338
hand pumping of, 175
hypereffective, 279
hypertrophy of, 319–320, 324
hypoeffective, 279
in cardiac output regulation, 272–274, 276, 279
in exercise, 276
in shock, 328, 331
lactic acid use by, 816
left, 150
failure of, 309, 312, 320
overloading of, 310
parasympathetic control of, 160–161, 170, 238, 691–692
right, 150, 287
failure of, 309–310, 312
sympathetic control of, 238, 691
temperature effect on, 162
Heart attack. See *Cardiac failure.*
Heart block, 171
abnormal rhythms from, 198–199
bundle branch, 190–191
Heart block (*Continued*)
complete, 199
incomplete, 198–199
of Purkinje system, 192
Heart failure See *Cardiac failure.*
Heart murmurs, 318–319, 323
Heart muscle See *Cardiac muscle.*
Heart rate, and duration of contraction, 153
atrial reflex control of, 250
effect of thyroid hormone on, 902
in cardiac pumping, 158, 160
in electrocardiogram, 179
in exercise, 1016
parasympathetic control of, 238
vasomotor center control of, 239
Heart sounds, 156–157, 316–319
Heart valves function of, 155–156
lesions of, 318–321, 324
scarring of, 318
Heartbeat, strength of, 902
Heartburn, 600
Heart-lung machines, 324
Heart lung preparation, 162–163
Heat. See also *Temperature*
acclimatization to, 859
body, during exercise, 1017
loss of, 850–852
production of, 850, 854–856
in energy release, 844
maintenance, 130
transfer to body core, 850
Heat stroke, 859, 1017
Helicotrema, 736, 737
Helium, in deep sea diving, 539, 641
Helium dilution method, 472
Helper cells, 83, 66
Hematocrit, 206–208, 384
at high altitudes, 531
Hemianopsia, bitemporal, 725
homonymous, 723
Hemiballismus, 629
Hemoglobin, adult, 498
carbon dioxide with, 501
carbon monoxide with, 500
concentration of, 42
effect of interstitial fluid oxygen pressure, 495
destruction of, 48
fetal, 498, 986
formation of, 46
in cyanosis, 522
oxygen binding with, 46, 496–497, 501–502
at high altitudes, 529, 530–531
oxygen-buffering function of, 5, 497–498
per cent saturation of, 496
transport of, 42
Hemoglobin dissociation curve, 496, 498–499
Hemoglobin S, 48–49
Hemolysis, 48–49
in transfusion reactions, 71–72, 74
Hemophilia, 80, 83
Hemopoiesis, by spleen, 343
Hemorrhage, from hypertension, 270–271
in thrombocytopenia, 83–84
shock caused by, 327–332, 333, 335
thirst caused by, 431
Hemosiderin, 47
Hemostasis, 76–77
Henderson-Hasselbalch equation, 440
Henry's law, 482
Heparin, 58, 82, 85
Heparinase, 85
Hepatic. See also *Liver.*
Hepatic arterioles, 835
Hepatic artery, 340, 341
Hepatic cellular plates, 835
Hepatic coma, 837
Hepatic congestion, 836
Hepatic duct, 781
Hepatic sinuses, 340
permeability of, 341
Hepatic sinusoids, 835
Hepatic vascular system, 836
Hering Breuer inflation reflex, 505
Hering's nerve, 247
Hermaphroditism, 1003
Herpes zoster, 601
Hexagonal reference system, 184
Hexamethonium, 283, 696
Hexokinase, 810
Hexosemonophosphate shunt, 48
hGH (human growth hormone), 891
High altitude, and respiratory alkalosis, 448
and tissue oxygenation, 43
effect on respiratory control mechanisms, 510
oxygen pressure at, 528–532
radiation at, 535
Hilus, 55
Hippocampus, 676, 682
in learning, 682–683
in memory, 660–661
in rage, 681
Hirschsprung's disease, 802
Histamine, as vasodilator agent, 242
in allergic reaction, 68
in anaphylactic shock, 332–333
in asthma, 521
in control of bronchioles, 475
in gastric secretion, 776–777
Histiocytes, tissue, 55
Histones, 36
Hives, 68
HLA antigens, 75
HMW kinogen, 80
Holistic theory, of thoughts, 657
Homatropine, 696
Homeostasis, 3–5
Horizontal cells, 720, 721
Hormone(s), blood concentrations of, 878–879, 882–883
chemistry of, 878–879
female, 968–969
general, 876
in arterial pressure regulation, 253–255
in athletic performance, 1008–1009, 1018
in regulation, 4–5
in smooth muscle contraction, 145
inhibitory, 679, 885, 886
local, 876
male, 960–964
mechanisms of action of, 879–882
metabolic clearance rate of, 882–883
nature of, 876
overview of, 876–878
releasing of, 679, 885, 886
secretion of, 878–879
rate of, 879, 883
storage of, 878–879
vs. factors, 886
Hormone receptors, 876, 879–880
Hormone-sensitive cellular lipase, 819
Horner's syndrome, 733

Human chorionic gonadotropin, 963
during pregnancy, 989
effect on corpus luteum, 972, 987-988
effect on fetal testes, 963
effect on fetal ovaries, 969
for lack of ovulation, 981-982
Human chorionic somatomammotropin, 877, 989, 991
Human chorionic thyrotropin, 989
Human growth hormone (hGH), 891
Humidification, of air, 484
Humoral regulation, of circulation, 241-243
Hunger, 759, 863
glucostatic theory of, 861
hypothalamic regulation of 679
Hunger center, 863
Hunger contractions, 763, 863
Huntington's chorea, 629
Hyaline membrane disease 467, 521, 1000
Hyaluronic acid, 21, 911
Hyaluronidase, 953, 958, 983
Hydrocephalus, 376-377
Hydrochloric acid secretion of, 774-775
Hydrocortisone See Cortisol
Hydrogen, in citric acid cycle, 812-813
in fat synthesis 816
ionization of, 813-814
release of, by phosphogluconate pathway 816
during glycolysis, 812
Hydrogen bonding, in DNA, 28
in proteins, 829
Hydrogen ion(s), countertransport of, 414
dissociation of carbonic acid into, 501
effect on oxygen hemoglobin dissociation curve, 498
in cerebral blood flow regulation, 339
in oxidative phosphorylation, 813-814
in respiratory control, 506-513
interaction with bicarbonate ions, 445
transport into urine 446-447
tubular secretion of, 406
active mechanism of, 401
aldosterone effect on, 912
by collecting ducts, 404
regulation of, 444-445
Hydrogen ion concentration, defense against changes in 439
effect on alveolar ventilation, 443-444
in acidosis and alkalosis, 438-439
renal regulation of, 444-448
tissue, 235
Hydrogen peroxide, in neutrophils and macrophages, 54
in thyroid hormone formation, 898
Hydrolases, 15-16
Hydrolysis, in digestion, 787
in small intestine, 788
of triglycerides, 818-819, 820
Hydrostatic pressure, 222-223
in bulk flow, 96-97, 351
in regional pulmonary blood flow, 289-291
Hydroxyapatites, 941-942, 952
β Hydroxybutyric acid, 821-822, 928
25-Hydroxycholecalciferol, 938-939
16-Hydroxydehydroepiandrosterone, 988
Hydroxyl ions, in neutrophils and macrophages, 54
in vascular control, 243
17-Hydroxyprogesterone, 973
17-Hydroxysteroids, 920, 921
5-Hydroxytryptamine See *Serotonin.*
Hyperadrenalism, 920-921
Hyperalgesia, 593, 601
Hyperalimentation, 392
Hyperbaric pressure chamber, 524
Hyperbarism, 537
Hyperbilirubinemia, physiological, 1002
Hypercalcemia, 940, 949
Hypercapnia, 516, 523
in oxygen therapy, 524
Hypercholesterolemia, familial, 827
Hypercomplex cells, 726, 727
Hypereffective heart, 161
Hyperemia, active, 234
reactive, 234
in coronary system, 297-298
in skin circulation, 345
Hyperestrinism, 973
Hyperglycemia, in giantism, 891
Hypergonadism, male, 965-966
Hyperinsulinism, 936
Hyperkalemia, 912
Hypermetropia, 705-706
Hyperopia, 705-706
Hyperosmolality, of medullary interstitial fluid, *414-416*
Hyperparathyroidism, 949
Hyperphagia, 863
Hyperplasia, 133
fibromuscular, 453
Hyperpnea, 516
Hyperpolarization, 116
after action potential, 659
at neuromuscular junction, 145
in familial periodic paralysis, 134
in gastrointestinal smooth muscle, 756
in inhibitory postsynaptic potential, 556
in sinoatrial fiber, 166, 170
Hyperpyrexia, 859
Hypertension, atherosclerosis with, 827
basic causes of, 264
defined, 264
effects of, 270-271
essential 269-270
Goldblatt, 268-269
in toxemia of pregnancy, 269
malignant, 271
neurogenic, 269
pulmonary, 520
renal, 457
spontaneous, 269
vasoconstrictor, 266-269
volume-loading, 264-266, 268-269
Hyperthyroidism, 905-906
cardiac failure from, 310
cardiac output in, 278
Hypertonicity, 390, 391
Hypertrophy, cardiac, 189-190, 191-192
from exercise, 1016
in congenital heart disease, 324
in valvular heart disease, 319-320, 324
muscle, 133-134, 1013
Hypoadrenalism, 920
Hypocalcemia, 940
Hypocapnia, 516
Hypochlorhydria, 799
Hypoeffective heart, 161
Hypogastric ganglia, 757
Hypogastric nerves, in labor pains, 993
Hypoglycemia, 936
Hypoglycemic shock, 926
Hypogonadism, female, 979-980
male, 965
Hypokalemia, 912
Hypoparathyroidism, 948-949
Hypophosphatemia, congenital, 950
renal, 458
Hypophysial stalk, 884, 893
Hypophysis See *Pituitary gland.*
Hypopnea, 516
Hypotension, arterial, 412-413
renal, 457
Hypothalamic inhibitory hormones, 885, 886
Hypothalamic releasing hormones, 885, 886
Hypothalamic thermostat, 853-857, 858
Hypothalamic-hypophysial portal system, 885-887
Hypothalamus, and limbic system, 676-678
behavioral functions of, 679-681
diffuse stimulation of, 240-241
endocrine functions of, 678-679
in ACTH secretion, 917-918
in female monthly rhythm, 977-978
in food intake regulation, 863
in growth hormone secretion, 890-891
in obesity, 865
in prolactin secretion, 995
in shivering, 855
in temperature regulation, 241, 344, 853-854, 858
in TSH secretion, 903-904
in vasoconstriction, 239
in vasodilatation, 240
pituitary gland and, 884-887, 963-964
posterior, 887-895
vegetative functions of, 678-679
Hypothermia, artificial, 860
Hypothyroidism, 906-908
atherosclerosis with, 827
Hypotonia, 647
Hypotonicity, 390, 391
Hypoventilation, alveolar, 518
hypercapnia from, 523
Hypovolemia, shock caused by, 327-332, 333, 335
Hypoxia, 516, 522-523
anaerobic energy during, 812
and erythropoietin formation, 44
at birth, 999
at high altitudes, 529-530
Cheyne-Stokes breathing in, 515
effect on synaptic transmission, 561
oxygen therapy for, 523-524
types of, 522-524

I (isotropic) bands, 121
IgE, 64, 68, 521
IgG, 64, 71
Iggo dome receptor, *581*
IgM, 64, 71
Ileocecal valve, 766-767
Ileum See also *Small intestine*
absorption in, 793
Immune reactions, in transplantation, 74-75
of blood, 70
Immune response, primary vs secondary, 63
Rh, 72-73
Immune system, macrophages in, 54
of infant, 1003
suppression of, 75
Immune tolerance, 67
Immunity, 60-67
effect of cortisol on, 917

Immunity (*Continued*)
aconatal, 1003
Immunoglobulins, 63-64
in allergies, 68
Impedance matching, 734-735
Implantation, of zygote, 984-985
Imuran, 75
Inactivation gate, 107
Inanition, 566
Inappropriate ADH secretion, syndrome of, 427, 431
Incisura, 156
at end of systole, 225
Incus, 734
Indicator dilution method, 284-285
Indocyanine green, 341, 836
Inducer substance, 34
Inductions, in embryogenesis, 38
Infant. See also *Neonate*
endocrine problems in, 1003-1004
immunity in, 1003
nutrition of, 1003
premature, 1004-1005
Infarct, anterior wall, 194
myocardial. See *Myocardial infarction.*
posterior wall, 194
rupture of, 301
Infection. See also *Inflammation*
cerebrospinal fluid pressure in, 376
pus formation in, 57
resistance to, 51-59
Infenur rollicular commissure, 741
Infenor obliques, 728-729
Infenor olivary nuclei, 633
Infenor recti, 728-729
Infertility, female, 981-983
male, 957-958
Infinite gain principle, 258-259
Inflammation, complement system in, 65
cortisol in, 916-917
macrophage and neutrophil response to, 56-57
Inflammation, 562
Information processing, 547-548
Ingestion, by cells, 19-20
of food, 759-762
Inhibin, in females, 972, 977
in males, 956, 964
Inhibition, allosteric, 815
lateral, 566-567
in dorsal column-lemniscal system, 587-588
in visual perception, 721, 722
of neuronal pool, 564
postsynaptic, electrical events in, 556
receptors in, 550, 552
presynaptic, 556-557
in primary memory, 658
reciprocal, 615
recurrent, 607
surround, 588
Inhibition circuit, lateral, 566-567
reciprocal, 565, 614
Inhibitor substance, 34
Inhibitory hormones, 679, 885, 886
Inhibitory neuron, 565
Inhibitory postsynaptic potential (IPSP), 556
summation of, 558
time course of, 557
Inhibitory receptors, 550, 552
in smooth muscle, 144, 145
Inhibitory state, 559
Inhibition zone, 564
Injury current, 192-195, 301
Innervation, reciprocal, 615
Inositol, as second messenger, 881
in phospholipid formation, 825
Inspiration, control of, 504-506
mechanics of, 466-467
work of, 469
Inspiratory capacity, 470
Inspiratory ramp, 505, 506
Inspiratory reserve volume, 470, 171
Insulator system, 849
Insulin, 887
and growth hormone, 888-889, 929
chemistry of, 923-924
for diabetes, 935
in carbohydrate metabolism, 924-926
in fat metabolism, 926-928, 930
in fat synthesis, 823
in glucose transport, 810
in protein metabolism, 834, 928-929
secretion of, 923
absence of, 927-928
control of, 929-930
effect of somatostatin on, 932
Insulin control system, 9-10
Insulin preprohormone, 924
Insulin shock, 936
Intellectual functions, analytical operations, 661-662
communication, 662
of specific cortical areas, 652-656
thought, consciousness, and memory, 656-661
transfer of information to opposite hemisphere, 662-661
Intensity, of neural signal, 562-563
of stimulus, 574, 576-577
Intention tremor, 645, 647
Interatrial septal defect, 321
Intercalated discs, 150
Intercellular cleft, 349
Intercostals, external, 466
internal, 467
Interference patterns, 727, 731
Interleukin, 63, 66
Intermediolateral horn, 687, 732
Internal environment, 3-5, 387
Internal sphincter, 460
Interneurons, 607
Internodal pathways, 165, 167
Interphase, 18, 35
Interpositus nucleus, 641, 644-646
Interstitial cells of Leydig, 960
tumors of, 965
Interstitial *fibrosis*, 518
Interstitial fluid, and interstitium, 351-352
blood exchange and, 83, 352-359
calcium in, 939-940
carbon dioxide pressure in, 496
colloid osmotic pressure of, 353, 357
composition of, 3
gel-matrix in, 370-371
medullary, hyperosmolality of, 414-416
of brain, 377
oxygen diffusion to, 494-495
protein concentration in, 363
control of, 365
in edema safety factor, 368
pulmonary, 371-372
volume of, 388, 428
Interstitial fluid compartment, 383
Interstitial fluid pressure, 353
control of, 365-366
in lymph flow, 363, 364, 365
in tightly enclosed tissues, 355
Interstitial fluid pressure (*Continued*)
measurement of, 354-355
negative, 364, 366
in edema safety factor, 368
in lungs, 292-293
of kidney, 396
positive, 367-368
pulmonary, 371-372
subatmospheric, 356
Interstitial nucleus, 626
Interstitial spaces, "dry" state of, 366
flow of fluids through, 3, 358
gel in, 370-371
pressure-volume curve of, 366-367
Interstitiospinal tract, 620
Interstitium, 351-352
Interventricular septal defect, 321, 323
Intestinal See also *Gastrointestinal*
Intestinal lipase, 785
Intestines See *Large intestine, Small intestine*
Intra-alveolar pressure, 467
Intracellular canaliculi, 774
Intracellular fluid, components of, 3, 88, 386-387
defined, 2, 88
nutrients in, 88
phosphate buffer in, 441
shifts between extracellular fluid and, 387-392
vs extracellular fluid, 3, 88
Intracellular fluid compartment, 383
Intrafusal muscle fibers, 607, 608
Intragel pressure, 354
Intralaminar nuclei, 590
Intramural plexus, 691
Intramyocardial pressure, 296
Intraocular fluid, 377-379
Intraocular pressure, 378-379
Intrapleural pressure, 467
Intrarenal pressure, 396
Intraventricular block, incomplete, 199
Intrinsic factor, deficiency of, 869
secretion of, 774, 775
vitamin B_{12} and, 45, 799
Inulin, tubular transport of, 405, 406
Inulin clearance, 407, 408
Involuntory fixation mechanism, 729-730
Involution, of corpus luteum, 972
of uterus, 994
Iodide ion, oxidation of, 898
Iodide pump, 898
Iodide trapping, 898
decreased, 904-905
Iodides, and thyroid activity, 905
ingestion of, 897
Iodinase, 898
Iodine, in formation of thyroxine, 897-898
ingestion of, 897
metabolic functions of, 872
Iodopsin, 712, 713
Ion(s), absorption of, 792-793, 796
and lipid bilayer, 90
hydrated, 90
in saliva, 772-773
osmotic effect of, 388
tubular transport of, 405-406
Ion *channels. See Protein channels.*
Ion gating, 91-92
Iothalamate, radioactive, 407
IPSP (inhibitory postsynaptic potential) 556
summation of, 558
time course of, 557

Iron, absorption of, 47, 793
daily loss of, 47
during pregnancy, 990
fetal metabolism of, 998
metabolism of, 46-48, 872
quantity in body, 46
regulation of, 47-48
required by infant, 1003
storage of 46-47, 838
transport of, 46-47
Iron sulfide proteins, 814
Irritable colon syndrome, 802, 803
Irritant receptors, of airways, 475, 513
Ischemia, as cause of pain, 591, 598
in Raynaud's disease, 315
myocardial, in aortic valvular disease, 320
ventricular fibrillation from, 174
of cardiac muscle, 194-195, 196
of vasomotor center, 250-251
renal 453
Ischemic heart disease, 299-303
Ischemic response, CNS, 250-251
in shock, 328
oscillation of, 233
Islets of Langerhans, 923
hormones of, 877
Isocitrate ions 926
Isografts, 74
Isogravimetric method, for measuring capillary pressure, 353
Isohydric principle, 442
Isomaltase, 785
Isometric contraction, 155, 159
Isometric relaxation, 155, 159
Isopropyl norepinephrine, 690
Isoproterenol, 696
Isosmotic absorption, 792
Isosthenuria, 454
Isotonicity, 390, 391
Isovolumetric method, for measuring capillary pressure, 354
Isovolumic contraction, 155, 159
Isovolumic relaxation, 155, 159
Itch 582

J point, 193
'J' receptors, 513-514
Jacksonian epilepsy, 675
Jaundice, 839-840
hemolytic, 74, 839-840
neonatal 1002
obstructive, 839, 840
Jejunum See also *Small intestine*
ionic absorption in, 793
Joint cavities, 381
Joint receptors, 588-589
of neck 625
Joint rotation, 588-589
Junctional feet, 127
Junctional fibers, 167
Junctional potential, 144-145
Juxtaglomerular cells, 412
erythropoietin formation by, 44
renin secretion by, 251
Juxtaglomerular complex, 403, 411-412
in blood flow control, 235

Kallidin, 242, 342
Kallikrein, 242, 345
Keratoconus, 707
Kernicterus, 73
Ketogenesis, 833
Ketogenic effect, of corticotropin and glucocorticoids, 824
of cortisol, 915-916
of insulin lack, 927-928
α-Ketoglutaric acid, 833
Ketone bodies, 822, 923
Ketosis, 822
from growth hormone, 888
from insulin lack, 928
in starvation, 866
17-Ketosteroids, 922
Kidney(s) See also *Renal, Tubular*
acid-base regulation by, 444-450
anatomy of, 393-394
angiotensin effect on, 262
artificial, dialysis with, 455-457
hypertension with, 266
atrial reflexes to, 250
blood flow through, 395-396
and glomerular filtration rate, 399
control of, 235, 413
measurement of, 408
clearance by, 458-459
concentrating mechanism of, 414-417
countercurrent mechanism of, 414-417
1,25-dihydroxycholecalciferol formation in, 938
diluting mechanism of, 413-414
diseases of, 452-458
hypertensive, 457
vascular, 453-454
erythropoietin formation by, 44
excretion by. See *Excretion*
fetal, 998
fluid retention by. See *Fluid retention*
fluid volume regulation by, 421-424
glomerular filtration in. See *Glomerular filtration.*
in arterial pressure control, 257-262
in cardiac failure, 306-307, 308, 309
in homeostasis, 4
in hypertension, 267-268, 457
in newborns, 1002
in shock, 334
pain in, 601
reabsorption by. See *Reabsorption*
salt retention by, 262-263
in cardiac failure, 306
secretion by. See *Tubular secretion*
Kidney disease, 452-458
hypertensive, 457
vascular, 453-454
Kidney failure, 452-457
Kidney function, in cardiac failure, 306-309
in newborns, 1002
in shock, 334
tests of, 458-459
Kidney shutdown, after transfusion reaction, 74
Kidney stones, in hyperparathyroidism, 919
Killer cells, 66
Kilogram-calorie, 844-845
Kinesiology, 133
Kinesthesia, 588, 589
Kinetochore, 37
Kinins, 242
in gastrointestinal blood flow, 342
Kinocilium, 622
Klüver Bucy syndrome, 682
Knee jerk, 612
Korotkoff sounds, 245
Krebs cycle, 23, 812-813
Kupffer cells, 55, 835, 836
in blood cleansing, 341
Kussmaul respiration, 934
Kwashiorkor, 862, 890

"Labeled line" principle, 573
Labor, 992-994
Labor contractions, 992-993
Labor pains, 993-994
Lactase, 785, 788
Lactation, 994-996
Lacteal, central, 791
Lactic acid, anaerobic glycolysis, 815-816, 1010, 1011-1012
in shock, 330
Lactic dehydrogenase, zinc in, 872
Lactose, 787
Lactotropes, 885
Laminar blood flow, 209-210
Langendorf preparation, 162
Langerhans' cells, 985
Language function 655
Laplace, law of, 214, 297
Large intestine See also *Colon.*
absorption in, 796
disorders of, 802-803
obstruction of, 805
secretions of, 785, 793
Larynx, 478
Late distal tubules, 403
osmolality in, 417
potassium secretion in, 420
sodium reabsorption in, 419
Lateral corticospinal tracts, 635
Lateral geniculate body, 724, 725
Lateral hypothalamus, behavioral functions of, 680, 681
in food intake regulation, 863
vegetative functions of, 678-679
Lateral inhibition, 566-567
in dorsal column-lemniscal system, 587-588
in visual perception, 721, 722
Lateral lemniscus, 740-741
Lateral motor system, 636
Lateral olfactory area, 751
Lateral recti, 728
Lateral spinothalamic tract, 590
Lateral vestibulospinal tract, 620
LATS (long acting thyroid stimulator), 905
Lead(s), electrocardiographic, 180-182
axis of, 184
increased voltage of, 191
potentials in, 184-185
Learning, hippocampus in, 682-683
reward and punishment in, 681
Lecithin, 782, 824
Left atrial pressure, 288, 291
Lemniscal system See *Dorsal column-lemniscal system.*
Lengthening reaction, 613
Lens(es), concave, 701
contact, 707
convex, 700-701, 702
corrective, 706-707
cylindrical, 701-702, 706-707
depth of focus of, 705
focal length of, 702
formation of image by, 702
refractive power of, 702-703
refractive principles of, 700-702
spherical, 701-702
Leprosy, 543

Leukemia(s), 58–59
Leukocytes, concentrations of, 51–52
genesis of, 52
life span of, 52–53
properties of, 53–58
types of, 51
Leukocytosis, 56
Leukorrhea, 977
Lever systems, 133
Leydig cells, 960
tumors of, 965
LH. See *Luteinizing hormone (LH)*.
Libido, hypothalamus in, 680
neuronal stimulus for, 959
sex hormones in, 684
thyroid hormone in, 903
Ligand gating, 91–92, 137
Light, adaptation to, 716–717
flickering, fusion of, 717
intensity changes of, 722–723
refraction of, 700
ultraviolet, 603
white, 718
Light reflex, pupillary, 733
Limbic cortex, 676–677, 683
Limbic system, amygdala, 681–682
behavioral functions of, 679–681
functional anatomy of, 676–677
hippocampus, 682–683
hypothalamus, 677–681
limbic cortex, 683
major output pathway of, 677–681
regulation of food intake by, 683
reward and punishment function of, 680–681
Liminal zone, 564
Linear acceleration, 623
Lines, detection of, 725–726
Lipases, hormone-sensitive cellular, 819
hormone sensitive triglyceride, 824, 927
in phagocytosis, 54
intestinal, 785
lipoprotein, 818–819
tissue, 820
Lipid(s). See also *Fat(s), Triglycerides*
dissolving of, 20–21
in cell, 12
in liver, 820, 901
plasma, 927
synthesis of, 21
transport in blood, 818–819
types of, 818
Lipid barrier (lipid bilayer), of cell membrane, 12–13
diffusion through, 901
protein transport through, 88–89
Lipid metabolism, and atherosclerosis, 826–827
and carbohydrate metabolism, 930
cholesterol in, 825–826
cortisol in, 915–916
fat deposits in, 820
growth hormone in, 888
in cell, 22
insulin in, 926–928
liver in, 837
phospholipids in, 824–825
rate of, 862–863
thyroid hormone in, 901
triglycerides in, 820–824
Lipid *nephrosis*, 455–456
Lipoprotein, 795, 819
Lipoprotein lipase, 818–819
Lipostatic theory, 864
Lipotropin, 597, 919
Liver. See also *Hepatic*
acetoacetic acid formation in, 821–822
amino acid synthesis by, 838
bile secretion by, 770, 781, 838–840
blood flow through, 430–432, 836
blood storage in, 836
blood-cleansing function of, 341
cirrhosis of, 342, 836
detoxification function of, 838
fatty, 927
glucose buffer function of, 837
glycogen storage in, 810–811, 924–926
in bleeding disorders, 83
in blood coagulation, 838
in newborns, 1002
in shock, 330
insulin effect on, 924–926, 927
intercellular clefts in, 349
iron storage by, 838
lipids in, 820, 901
lipoprotein formation in, 819
lobules of, 835
lymph flow from, 836
macrophages in, 55, 835, 836
metabolic functions of, 836–838
nutrients supplied by, 4
physiologic anatomy of, 835–836
proteins in, 915
reservoir function of, 341
tissue macrophage system of, 836
urea formation by, 833, 837
vitamin storage by, 838
Load, and velocity of contraction, 126
in isotonic contraction, 130
Load reflex, 610
"Lochia," 994
Locomotion, ameboid, 24–25
ciliary, 25–26
Locus ceruleus, in depression, 683
in REM sleep, 673
in reticular activating system, 667–668
norepinephrine release by, 553
Long acting thyroid stimulator (LATS), 905
Loop of Henle, 393, 400
as countercurrent "multiplier," 416
ascending limb of, 393–394, 403, 417
descending limb of, 393, 403
osmotal concentration in, 417
potassium reabsorption by, 420
sodium reabsorption by, 419
thin and thick segments of, 403
Loudness, 739–740
LRH (luteinizing hormone releasing hormone), 886
in male sexual function, 963–964
in ovarian cycle, 968, 977
Lubrication, female, 980
male, 959
Luminosity, 727
Lumirhodopsin, 714
Lung(s). See also *Pulmonary, Respiratory system*.
as blood reservoir, 289
at birth, 993
blood flow through, 289–291
obstruction of, 293–294
blood volume of, 288–289
capillary exchange of fluids in, 292–293, 372
carbon dioxide removal by, 4
compliance of, 468–469
constricted, 476
diffuse sclerosis of, 294
expansion and contraction of, 466–467
interstitial fluid dynamics in, 371–372
Lung(s) (*Continued*)
massive collapse of, 520–521
negative interstitial fluid pressure in, 292–293
removal of one, 293
Lung capacity, total, 471
Lung disease, chronic obstructive, 491–492
Lung recoil pressure, 467
Lung volumes, 517
Lupus erythematosus, 67
Lutein cell, 971
Luteinization, 971
Luteinization inhibiting factor, 972
Luteinizing hormone (LH), 877, 884
in male sexual function, 963–964
in menopause, 979
in ovulatory cycle, 968, 969–972, 977–978
in spermatogenesis, 956
Luteinizing hormone releasing hormone (LRH), 886
in male sexual function, 963–964
in ovarian cycle, 968, 977
Lymph, chylomicron transport in, 796
flow of, 363–365
as edema safety factor, 368
from liver, 836
from lungs, 371
formation of, 363
Lymph channels, 361
Lymph duct, 361
Lymph nodes, 363
macrophages of, 55
organization of, 55
Lymphatic capillaries. See *Lymphatics*.
Lymphatic capillary pump, 364
Lymphatic pump, 363–364, 365
in drainage of potential spaces, 380
Lymphatic system, 361–365
in potential spaces, 379–381
perivascular spaces as, 375–376
Lymphatics, 361
afferent, 55
collecting, 363, 364
efferent, 55
external compression of, 364
intrinsic pumping by, 364
obstruction of, 369, 380
pulmonary, 287–288
Lymphoblasts, 63
Lymphocytes, 51
activated, 60
B, 61–65
clones of, 62–63
cortisol effect on, 917
genesis of, 52
life span of, 52–53
T. See *T cell(s)*
Lymphoid tissue, 61
lymphocyte spread to, 62
T cell release from, 65–66
Lymphokines, 63, 66
Lysine, and insulin secretion, 930
Lysine monohydrochloride, for alkalosis, 450
Lysis, of antigen, 64, 65
of blood clots, 82–83
Lysoferrin, 20
Lysosomes, cell, 15–16
digestion of foreign substances by, 20–21, 54
formation of, 22
stabilization by cortisol, 916–917
Lysozyme, in cell, 20
in immunity, 60

Lysozyme (*Continued*)
in infection, 54
in saliva, 773
Lytic complex, 72

Machinery murmur, 323
Macrocyte, 45
Macromotor units, 131
Macrophage(s) activation of, 66
alveolar, 55 477
fixed, 55
in immune system, 54
in lymphocyte activation, 63
phagocytosis by, 54
properties of, 53–54
response to inflammation, 56, 57
tissue, 52, 54–56
Macrophage migration inhibition factor, 66
Macula, retinal, 711
vestibular, 621–622, 623
Macula densa, 411
Magnesium, in vascular control, 243
metabolic functions of, 871–872
Magnesium ion concentration, *regulation of*, 436
Magnet reaction, 615
Magnocellular portion, of red nucleus, 636
Malleus, 734
Malnutrition, effect on metabolic rate, 847
Malocclusion, 952, 953
Malonyl CoA, 822
Maltase, 785, 788
Maltose, 787, 788
Maltotriose, 787
Mammillary bodies, 679
Manic-depressive psychoses, norepinephrine and serotonin systems in, 684
Mannitol, 407 450
Manometer, 211
Margination, of leukocytes, 53, 56, 57
Mark time reflex, 616
Mass action effect, 240, 815
Mass discharge, 694–695
Mass movements, of small intestine, 767–768
Mass reflex, 617
Mast cells, 58, 65
heparin release by, 82
Mastication, 759–760
Maturation, of ovum, 983
of sperm, 956
Maturation factor, in red blood cell formation, 45
Maturation failure, in red blood cell production, 45
Maximum breathing capacity, in deep sea diving, 541
Maximum expiratory flow, 476–477
Mean arterial pressure, 241–245
Mean circulatory filling pressure, 215
and blood reservoirs, 225
during exercise, 252
Mean electrical axis, of ventricles, 188–191
Mean pulmonary filling pressure, in left heart failure, 309
Mean QRS vector, 188–191
Mean systemic filling pressure, after cardiac failure, 306, 308
in cardiac output regulation, 276
ous return, 251
Mechanoreceptive sensation, detection of, 580–581
thalamus in, 590
transmission of, in anterolateral system, 582, 583, 589–590
in dorsal column–lemniscal system, 582–589
in peripheral nerve fibers, 581–582
Mechanoreceptors, 572
adaptation of, 575–576
Meconium, 998
Medial forebrain bundle, 677
in pain control system, 597
in reward, 680
Medial geniculate nucleus, 741
Medial lemniscus, 582, 583, 584
Medial longitudinal fasciculus, in eye movements, 729
in stereotyped body movements, 626
in vestibular system, 623
Medial motor system, 620, 636
Medial olfactory area, 751
Medial recti, 728
Medial vestibulospinal tract, 620
Median eminence, 679, 885–886
Medulla, adrenal, 692–693, 909
in vasoconstriction, 240
neural stimulation of, 687, 876
autonomic control by, 695
pyramids of, 635
renal, 393, 414–416
Medullary reticulospinal tract 620
Medullary sinuses, 55
Megacolon, 802
Megaesophagus, 798
Megakaryocytes 51
genesis of, 52
platelet formation from, 76
Megaloblasts, 45, 48
Meiosis, 955
Meissner's corpuscle, 580–581
Meissner's plexus, 756–757
Melanin, in Addison's disease, 920
in retina, 712–713
secreted with ACTH, 919
Melanocyte-stimulating hormone (MSH), 919, 920
Melanocytes, 919
Melatonin, 966
Membrane, nuclear, 11, 18
of cell See *Cell membrane*
of mitochondrion, 17
permeability of, 93
postsynaptic, in memory, 659
receptors on, 551–552
synaptic excitation of, 555
replenishment of, 22
respiratory, 486–490, 518
selectively permeable, 95
"short circuiting" of, 556
Membrane channels See *Protein channels*
Membrane potential(s) See also *Action potential(s)*
defined, 101
from active transport, 103, 104–105
from diffusion, 101–103
in smooth muscle, 141
measurement of, 103–104
physics of, 101–104
recording of, 117–118
resting, 104–106
in cardiac muscle, 151
in skeletal muscle, 127
of neuronal soma, 553–555
of sinoatrial fiber, 166
slow wave rhythm of, 142–143
Membrane stabilizing factors, 117
Membranous labyrinth, 621
Memory, 656–657
consolidation of, 660, 683
physiological basis of, 658–661
reward and punishment in, 681
searching for, 669
storage of, 548
transfer to opposite hemisphere, 662–663
types of, 657–658
Memory cells, 63, 66
Menarche, 978–979
Meningitis, headache from, 602
Menopause, 979
Menorrhagia, 903
Menstrual cycle See *Ovarian cycle*
Menstruation, 976–977
irregular, 979–980
prevention of, 987–988
Mercuric ion, 453
Meridional fibers, 690, 704
Merkel's discs, 581
Mesencephalon, autonomic control by, 695
Mesenchymal tissue, 985
Mesenteric ganglia, 757
Mesodermal axis, of embryo, 38
Mesolimbic dopaminergic system, 684
Messenger RNA, 30, 31–32
Mestranol, 981
Metabolator, 845
Metabolic clearance rate, of hormones, 882–883
Metabolic mass to tissue vascularity proportionality, 237
Metabolic rate, 841–848
basal, 847–848
during pregnancy, 991
effect of estrogens on, 974
effect of testosterone on, 962
effect of thyroid hormones on, 900–901
neonatal, 1003
Metabolism, ATP in, 23–24, 808–809, 841–843
carbohydrate See *Carbohydrate metabolism*
during lactation, 995–996
during pregnancy, 990
effect on interstitial fluid carbon dioxide pressure, 496
effect on interstitial fluid oxygen pressure, 495
enzymes in, 843–844
fetal, 998–999
growth hormone effect on, 887–889
in action potential, 111–112
in cardiac output, 272, 274–275, 276
in coronary blood flow control, 296–298, 299
in cutaneous blood flow regulation, 345
in exercise, 276, 1009–1014
in hypovolemic shock, 333
in newborns, 1002–1003
lipid. See *Lipid metabolism*
mitochondria in, 22–24
of minerals, 871–872
of oxygen, 22, 499
protein See *Protein metabolism*
rate control of, 843–844
substrate in, 844
thyroid hormone in, 684, 901
vitamin, 867–871
effect of thyroid hormone on, 901
fetal, 998–999
Metaphase, 37
Metarhodopsin, 714

Metarterioles, 231, 348
Methacholine, 138, 696
Methoxamine, 696
Methyl mercaptan, 750
Methylprednisone, 910
Micelles, 782, 783
"ferrying" role in absorption, 789, 790, 795
formation of, 789
Microcirculation, 231-232
Microfilaments, 16
in mitosis, 37
Microglia, 55
Microtubules, 16
Microvilli, 746, 791-792
Micturition, 460-462
Micturition contraction, 461
Micturition reflex, 461-462
Micturition waves, 461
MIF (müllerian inhibitory factor), 956
Migraine headache, 603
Migrating myoelectric motor complex, 759
Milieu intérieur, 3
Milk ejection, effect of oxytocin on, 895, 995
regulation by hypothalamus, 679
Milk letdown, 895, 995
Mineral metabolism, 871-872
Mineralocorticoids, 909, 910
deficiency of, 920
functions of, 911-914
Minute respiratory volume, 472, 474
Minute ventricular output curve, 159
Miosis, 733
Mitochondrion(ia), 17-18
chemiosmotic mechanism of, 813-814
effect of thyroid hormone on, 900-901
energy extraction by, 22-24
entry of fatty acids into, 820
in muscle fiber, 122
in terminals, 550
Mitosis, 18, 35, 36-37
Mitotic apparatus, 36
Mitotic spindle, 16, 36
Mitral cells, 751
Mitral regurgitation, 319, 320-321, 324
Mitral stenosis, 319, 320-321, 324
Mitral valve, 155
lesions of, 319, 320-321
Mixing movements, 758
in colon, 767
in small intestine, 765
in stomach, 762
Modiolus, 736
Molar concentration, in osmotic pressure, 95
Monoamine oxidase, 689
Monoamine oxidase inhibitors, for depression, 684
Monocytes, 51
genesis of, 52
life span of, 52
properties of, 53-54
response to inflammation, 57
Monoglycerides, 789, 795
Monoiodotyrosine, 898, 899
Monosaccharides, 787, 788, 794
Monosynaptic pathway, 609-610
"Moon face," 921
Morula, 984, 989
Mossy fibers, 642
Motor aphasia, 662
Motor apraxia, 634
Motor area, *primary*, 632
excitation of spinal cord by, 636-638
removal of, 638
Motor area (*Continued*)
topographical map of, 633-638
supplemental, 633-634
Motor association area, 634
Motor axis, 547
Motor cortex, anatomy of, 632-633
effect of lesions in, 638
fiber pathways to and from, 634-635
in vasoconstriction, 240, 241
in vasodilation, 240
somatic sensory feedback to, 637
vertical columnar arrangement of neurons in, 636-637
Motor engram, of rapid motor patterns, 649
Motor functions, 547
cerebellar control of, 643-647
cortical control of, 632-638
of basal ganglia, 626-631
of brain stem, antigravity, 619-621
in subconscious, stereotyped movements, 625-626
vestibular sensations, 621-625
of spinal cord. See *Spinal cord reflexes.*
sensory feedback control of, 648-649
voluntary, initiation of, 649
Motor neuron, alpha, 607
anatomy of, 550
anterior, 607
gamma, 607
spinal, 637-638
Motor unit(s), 131, 607
asynchronous summation of, 132
Mountain sickness, 532
Mouth, digestion in, 787-788
dryness of, 431-432
Movement(s), ballistic, 645-648
cerebellar control of, 643-647
circus, 172-175
flexing, 626
hand, 634
of colon, 767-769
of eye, control of, 634, 728-729
fixation, 729-731
opticokinetic, 730
pursuit, 730-731
raising, 626
rapid motor patterns, 649
rotational, 626
sensory engrams of, 648
stepping, 616
stereotyped, 625-626
turning, 626
walking, 616, 626
Movement receptors, 576
Moving parallax, 708
MSH (melanocyte stimulating hormone), 919, 920
Mucin, 772
Mucopolysaccharides. See *Proteoglycans.*
Mucosa, 754
Mucous cells, gastrointestinal, 770, 784
Mucous glands, simple vs. compound, 774, 784
Mucous neck cells, 774
Mucus, importance of, 772
in large intestine, 785
in respiratory passageways, 475
in small intestine, 784
in stomach, 774, 775-776
Müllerian inhibitory factor (MIF), 956
Murmurs, 318-319, 323
Muscarinic drugs, 696
Muscarinic receptors, 689-690
Muscle(s), cardiac. See *Cardiac muscle.*
contraction of. See *Contraction(s).*
Muscle(s) (*Continued*)
effect of athletic training on, 1013
effect of insulin on glucose metabolism in, 925-926
effect of testosterone on, 962
effect of thyroid hormone on, 902
endurance of, 1009
glycogen storage in, 810-811, 925
in exercise, 1009-1014
metabolic systems of, 1009-1014
red, 131
skeletal. See *Skeletal muscle*
smooth. See *Smooth muscle.*
strength of, 1009
white, 131
Muscle control, muscle receptors in, 607-613
Muscle cramps, 598-599, 617
Muscle fatigue, 133, 1011
Muscle fiber, fast vs. slow twitch, 131, 1013-1014
skeletal, 120-122
excitation of, 127, 137-138
Muscle impulse, 111
Muscle jerks, 612
Muscle length, and spindle receptor excitation, 608-609
feedback control of, 610
"Muscle pump," 223
Muscle receptors, 607-613
Muscle spasm, as cause of headache, 603
as cause of pain, 594, 598-599
drugs causing, 139
from lesions of motor cortex, 638
spinal cord reflexes causing, 617
Muscle spindle, 133, 607, 608
continuous discharge of, 609
in voluntary motor activity, 611-612
innervation of, 608
receptor function of, 608-609
signal averaging function of, 810
structure of, 608
Muscle tension, active vs. resting, 128
detection of, 612-613
Muscle tone, inhibition by basal ganglia, 628
of skeletal muscle, 132-133
of smooth muscle, 146
Muscle tremor, effect of thyroid hormone in, 902
Muscle twitches, 130-131
Muscle weakness, from hypokalemia, 912
Muscular dystrophy, *swallowing disorders in, 798*
Myadriasis, 733
Myelin sheath, 114
degeneration of, 868
Myenteric plexus, 756-757, 758-759, 768
Myenteric reflex, 759
Myocardial infarction, 300
acute, death from, 300-301
heart function following, 302
recovery from, 301-302, 307
old recovered, 195
subendocardial, 300
Myocardial ischemia, in aortic valvular disease, 320
ventricular fibrillation from, 17
Myocardium, in decompensated failure, 308-309

Myocardium (*Continued*)
recovery after myocardial infarction, 307
Myoepithelial cells, 995
Myofibrils, 121-122
Myogenic mechanism, 413
Myoglobin, 46
Myopia, 706
Myosin in ameboid locomotion, 24
in mitosis, 37
molecule of, 123
Myosin cross bridges, 121, 123, 124-125
Myosin filaments, 121
characteristics of, 123-124
in muscle contraction, cardiac, 150
skeletal, 124-126
smooth 140, 141
Myosin head, 123-124
in skeletal muscle, 125
in smooth muscle, 143-144
Myotatic reflex, 609-610, 612
cerebellum in, 644
Myxedema 907

NAD (nicotinamide adenine dinucleotide), 812-814, 868
NADP (nicotinamide adenine dinucleotide phosphate), 816, 868
Na K ATPase, 901
Na K pump See *Sodium potassium pump*
Natriuresis, for essential hypertension, 270
from diuretics, 459
pressure, 257, 258-260
Natriuretic factor atrial, 427
Nausea, 804
Near-sightedness, 706
Neck, joint receptors of, 625
Neck reflexes, 625
Necrosis tissue in severe shock, 330
tubular, acute renal failure in, 452-453
from shock 334
Negative feedback, in control system, 6
in glomerular filtration rate, 412-413
in hormone secretion 879
of operon, 34
Neocerebellum, 639
Neocortex, 677
Neonate See also *Infant*
adjustments to extrauterine life, 999-1001
cyanosis in, 323-324, 522
nutrition of, 1001
special functional problems in, 1001-1004
Neostigmine, 139, 696
Neostriatum, 627
damage to, 629
function of, 628
Nephrons See also *Tubules*.
cortical, 394
function of, 395
in renal failure, 453-454
juxtamedullary, 394, 414
structure of, 394
Nephrosclerosis, benign, 453-454
Nephrosis, edema in, 369
lipoid, 457-458
Nephrotic syndrome, 457-458
Nernst equation, 94, 102, 554
Nernst potential, 101, 102
deafness, 743, 744

Nerve fiber(s), centrifugal, 566
excitability curve of, 116-117
"fast" and "slow" pain, 594
gamma motor. See *Gamma motor nerve fibers*.
heat production by, 112
myelinated, 114
physiological classification of, 578-579
"recharging" of, 111-112
saltatory conduction in, 114-115
signal transmission in 114-115, 562-563
transmission of tactile sensations in, 581-582
unmyelinated, 114
"Nerve" gas, 139
Nerve impulses, 111, 112, 562
and signal strength, 562-563
energy for conduction of, 841-842
rate of, 574-575
transduction of sensory stimuli into, 573-576
Nervi erigentes, 687
Nervous system See also *Autonomic nervous system*, *Central nervous system (CNS)*
effect of thiamine deficiency on, 867-868
enteric, 756-758, 770
fetal, 998
general design of 546-548
in arterial pressure regulation 247-253
in homeostasis, 4
information processing in, 547-548, 562
integrative portion of, 4, 547-548
memory and, 549
motor portion of 4 547
peripheral in alkalosis, 449
sensory portion of, 4, 546-547
stabilization of, 569-571
via electronic computer, 549
Neural adaptation, to light, 717
Neurogenic hypertension, 269
Neurogenic shock, 332, 335
Neurohypophysis, 884, 893-895
Neuromuscular junction, 136
drugs affecting, 138-139
fatigue of, 138
of skeletal muscle, 136-139
of smooth muscle, 144-145
physiologic anatomy of, 136-137
safety factor at, 138
Neuromuscular transmission, 136-139
Neuron, anatomy of, 550
dynamic, 637
excitatory state of, 559-560
facilitation of, 558, 564
centrifugal control of, 566
firing rate of, 559-560
inhibitory, 565
motor, alpha, 607
anatomy of, 550
anterior, 607
gamma, 607
spinal, 637-638
postganglionic, 687-688
drugs stimulating, 696-697
postsynaptic, 549, 551-552
preganglionic, 687-688
presynaptic, 549-551
static, 637
synaptic functions of, 549
Neuronal circuits, inhibitory, feedback, 570

Neuronal circuits (*Continued*)
gross control by, 570
lateral, 566-567
reciprocal, 565
instability and stability of, 569-571
oscillatory, 567-569
reverberatory, 567-569
Neuronal columns, 726
Neuronal excitation, electrical events during, 553-556
receptors in, 550, 552
role of dendrites in, 558-559
threshold for, 556, 559-560
Neuronal inhibition, electrical events in 556
receptors in, 550, 552
Neuronal pools, signal transmission in, 563-569
Neurophysins, 893
Neurotensin, 886
Neurotransmitters, at smooth muscle neuromuscular junction, 144
defined, 548
excitation by, 116, 551, 552
inhibition by, 551, 552
release of, 550-551, 553
removal of, 552
substances that function as, 552-553
synaptic modulators, 552
synthesis of, 551
Neutralization, of antigen, 61 65
Neutrophil(s), phagocytosis by, 54
polymorphonuclear, 51
properties of, 53-54
response to inflammation, 56-57
Neutrophilia, 56, 57
Niacin, 868
Nicotinamide adenine dinucleotide (NAD$^+$), 812-814, 868
Nicotinamide adenine dinucleotide phosphate (NADP$^+$), 816, 868
Nicotine 138
Nicotinic acid, 868
Nicotinic drugs, 696-697
Nicotinic receptors, 689, 690
Night blindness, 714
Nigrostriatal tract, in Parkinson's disease, 630
Ninth nerve, 687
Nitrates, reabsorption of, 405, 406
Nitrogen balance, 862
Nitrogen narcosis, 537
Nitrogen pressure, in deep sea diving, 537, 539-541
in entrapped air, 525
Nitrogen volume, in deep sea diving, 539-540
Nitroglycerin, 302
Nitrous oxide, in coronary blood flow measurement, 303-304
Nociceptive reflex, 613-614
Nociceptors, 572
Nocturnal emissions, 959
Nodal fibers, 167-168
Node of Ranvier, 114
Norepinephrine, as adrenergic drug, 696
as neurotransmitter, 553
at smooth muscle neuromuscular junction, 144
derivation of, 878
general nature of, 876
in action potential, 116
in anaphylactic shock, 335
in autonomic nervous system, 688-690, 692-693
in brain activation, 684

Norepinephrine (*Continued*)
in cardiac failure, 309
in cardiac impulse transmission, 171
in control of bronchioles, 475
in coronary blood flow, 298
in cutaneous blood flow regulation, 344-345
in depression, 683-684
in fat utilization, 824
in hypertension, 267
in muscle blood flow regulation, 337
in vasoconstrictor system, 238-241
arterial, 253
secretion of, 909
Norethindrone, 981
Norethynodrel, 981
Norgestrel, 981
Nose, in smell, 749-750
respiratory functions of, 477
Nuclear bag fibers, 608
Nuclear chain fibers, 608, 609
Nuclear envelope, 18
Nuclear membrane, 11, 18
Nuclear pores, 18
Nucleic acid, 19
Nucleolus(i), 18
ribosome formation in, 31
Nucleoproteins, 12
Nucleotides, DNA, 23
RNA, 29-30
Nucleus, 11
evolution of, 19
structure of, 18
Nucleus ambiguus, 503
Nucleus of the tractus solitarius, in respiration, 504
in sleep production, 672
in taste, 747
Nucleus parabrachialis, 505
Nucleus precommissuralis, 626
Nucleus retroambiguus, 505
Nutrient demand theory, 233-234
Nutrients, absorption of, 794-796
capillary diffusion of, 350-351
energy extraction from, 22-24
in exercise, 1012-1013
in extracellular fluid, 4
Nutrition, during pregnancy, 990
of embryo, 985
of infant, 1003
of newborn, 1001
Nutritional regulation, 863, 864
Nystagmus, cerebellar, 647
pathologic, 731
vestibular mechanism for, 625

O A-B blood groups, 70-72
Obesity, 824, 865-866
caused by cortisol, 916
hereditary, 824, 865
psychogenic, 865
Occlusion, 951
Odontoblasts, 951
Odor blindness, 750
Ohm's law, 208
Oleic acid, 818
Olfaction, 749-751
Oligomenorrhea, 903
Olivocerebellar fibers, 635
Olivocerebellar tract, 640
Oncogenes, 38
Oncotic pressure, 221, 356
On-off response, to light intensity, 722-723
Oocyte, 983
Oocyte maturation-inhibiting factor, 969
Opening shock load, 534
Operator, activator, 31
repressor, 33-34
Opercular insular area, 748
Operon, 33-34
Ophthalmoscope, 708-709
Opiates, in pain control system, 597
Opsins, 716
Opsonization, 20
by neutrophils and macrophages, 53-54
in complement cascade, 65
Optic chiasm, 724
destruction of, 728
Optic disc, 728
effect of high cerebrospinal fluid pressure on, 376
Optic nerve, destruction of 728
in visual pathway, 724
signals transmitted through, 722-723
Optic radiation, 724
destruction of, 728
Optic tracts, 724, 725
lesions of, 728
Optical instruments, 708-709
Opticokinetic movements, 730
Optics, of eye 703-708
physical principles of, 700-703
Orbitofrontal area, 676, 683
Organ of Corti, 735, 738-739
Organ transplantation, 74
Organelles, evaluation of, 19
of cytoplasm, 14-18
Orgasm, female, 980-981
male, 960
Orientation, of head, 621-622
of lines and borders, 726
Ornithine, 833
Oscillatory circuit, 567-569
Oscillometer, 245-246
Oscilloscope, cathode ray, 117-118
electrocardiogram recording with, 179
Osmolality, 95-96, 388
extracellular fluid, 428-433
in vascular control, 213
of body fluids, 388-389, 390-392
Osmolar clearance, 417
Osmolarity, 96, 388
Osmole, 95-96, 388
Osmoreceptors, 429-431, 894
Osmosis, 387-389
across selectively premeable membranes, 94-97
rate of, 388
through tubular epithelium, 401
Osmosodium receptor, 429-431
Osmotic diuresis, 422
Osmotic equilibrium, 389-390
Osmotic pressure, 95-96
colloid See *Colloid osmotic pressure*
total, 356, 389
Osmotic pressure gradient, 96-97, 351
Osmotic regulation, of ADH production, 893-895
Ossicular system, 734-735
Osteitis fibrosa cystica, 949
Osteoblasts, 941, 942, 943, 945
Osteoclasts, 942-943, 945
Osteocytes, 941
Osteocytic membrane system, 945
Osteoid, 941
Osteolysis, 945
Osteomalacia, 950
in renal failure, 455
Osteon, 943
Osteoporosis, 950, 974
Osteoprogenitor cells, 943, 945
Ostium, of fallopian tube, 983
Otoliths, 621
Otosclerosis, 744
Oval window, 736, 737
Ovarian cycle, 969-972
anovulatory, 978
fertile period of, 981
psychic factors in, 977
regulation of, 977-979
suppression during nursing, 995
thyroid hormone effect on, 903
Ovarian follicle, 968-970, 978
Ovarian hormones, 968
abnormalities in secretion of, 979-980
functions of, 972-977
interplay with hypothalamic pituitary hormones, 977-979
relaxin, 989-990
target tissues of, 876
Ovaries, 968
Overdrive suppression, 170, 199
Overflow dribbling, 462
Overshoot potential, 151
Ovulation, 970-971
lack of, 978, 981-982
suppression of, 981
Ovulatory surge 971
Ovum(a), fertilization of, 983-984
implantation of, 984-985
maturation of, 983
number of, 968
primordial, 968
transport of, 983 984
Oxalate ion, as anticoagulant, 81, 85
Oxaloacetic acid, 812
Oxidases 16
Oxidation, beta, 820-821
control of, 815
in energy release, 129-130
of acetyl CoA, 821
of deaminated amino acids, 833
of fatty acids, 820-821
Oxidative enzymes, 23
Oxidative phosphorylation, 813-814
Oxidizing agents, in neutrophils and macrophages, 54
Oxidizing free radicals, 538-539, 543
Oxygen, alveolar, 289
energy equivalent of, 845
hemoglobin binding of, 46, 496-497, 501-502
at different altitudes, 529, 530-531
in blood flow regulation, acute, 232-234
arterial, 250
long term, 236
pure, breathing of, 529, 542
recycling of, 535
Oxygen concentration, arterial, 508-509, 513
in alveoli, 485, 490-492
in extracellular fluid, 5
in regulation of vasomotion, 349
Oxygen consumption, in exercise, 1014, 1016
myocardial, 297-298
Oxygen debt, 843, 1011-1012
Oxygen deficiency, in alveolar ventilation, 509-511
in cerebral blood flow regulation, 339
in irreversible shock, 331
in respiratory control, 509-511
Oxygen demand theory, 233-234, 296-297
Oxygen diffusion, capacity for,

Oxygen diffusion (*Continued*)
from capillaries to cells, 495
from capillaries to interstitial fluid, 494-495
in athletes, 1014-1015
through placental membrane, 986-987
Oxygen, Fick method, 284
Oxygen poisoning, 500
in deep sea diving, 537-539
Oxygen pressure, 493-496
and rate of oxygen usage, 499
at high altitude, 528-532
determination of, 517
in deep sea diving, 537-539
in entrapped air, 525
in respiratory control 506, 508-511
Oxygen therapy, for hypoxia, 523-524
for shock, 334
hyperbaric, 543
in premature infant, 1004-1005
Oxygen transport, abnormalities of, 518
and red blood cell production, 43-44
effect of high oxygen pressure on, 537-538
in arterial blood, 494
in dissolved state, 499-500
in exercise, 494, 497
in pulmonary blood, 493-494
Oxygen utilization, at high altitudes, 531
by cell, 499
Oxygenators, 321
Oxygen-hemoglobin dissociation curve, 496, 498-499
Oxyntic cells, 774
Oxyntic glands, 774-775
Oxyphil cells, 944
Oxytocin, 877, 884
chemical nature of, 893
effect on uterus, 895, 992
fetal, 992
in female orgasm, 980
in milk ejection, 895, 995

P 4000, 795
P vectorcardiogram, 188
P waves, 154, 177 178, 187
Pacemaker, abnormal (ectopic), 169-170
artificial, 199
S-A node as, 169-170
Pacemaker waves, 143
Pacinian corpuscle, adaptation of, 575-576
in joints, 588, 589
receptor potential of, 574
stimulation of, 581
PAH (para aminohippuric acid), as measure of plasma flow, 408
transport by renal tubules, 405, 406
Pain, abnormalities of, 601-602
acute, 592, 594-595
detection of, 593-594
headache, 602-603
localization of, 596, 599-600
measurement of, 592-593
"parietal," 599, 600-601
purpose of, 592
referred, 598, 600
slow, 592, 594-596
thalamus in, 590
transmission of, 594-596
visceral, 598-601
Pain control system, 596-598
Pain inhibitory complex, 597
pathways, 594-596
Pain receptors, 593-594
in thermal sensation, 603-604
Pain reflex, 613-614
Pain threshold, 592-593
Paleocerebellum, 639
Paleocortex, 677
Palmitic acid, 818
Pancreas, hormones of, 877
pain in, 600
physiologic anatomy of, 923
secretions of, 770, 778-779
Pancreatic amylase, 779, 788
Pancreatic lipase, 779, 789
Pancreatic polypeptide, 923, 932
Pancreatitis, 779, 801
Panhypopituitarism, 891
Panting, 513, 853
Pantothenic acid, 870
Papillae, of Vater, 801
renal, 391
Papillary muscles, 156
Papilledema, 376
Para aminohippuric acid (PAH), as measure of plasma flow, 408
transport by renal tubules, 405, 406
Parachute jumps, 531
Parafollicular cells, 917
Parahippocampal gyrus, 676
Parallax, moving, 708
Paralysis, diver's, 540
familial periodic, 117, 131
from thiamine deficiency, 868
of respiratory muscles, 471
Paralysis agitans 629-631
Paraolfactory area, 676
Parasympathetic nervous system, bladder control by, 460-461
bronchiole control by, 475
cholinergic and adrenergic fibers in, 688-689
denervation supersensitivity with, 693-694
effects on specific organs, 690-692
excitatory and inhibitory effects of, 690
gastrointestinal innervation by, 757
heart control by, 160-161, 170, 238
in coital lubrication, 959, 980
in defecation, 768
in erection, 959, 980
in gastrointestinal secretion, 770-771
in pancreatic secretion, 780
in pupillary light reflex, 733
in salivary secretion, 773
mass stimulation by, 694-695
physiologic anatomy of, 687-688
receptors of effector organs in, 689-690
stimulus rate in, 693
Parasympathetic potentiators, 696
Parasympathetic tone, 693
Parasympathetic transmitters, 688
Parasympathomimetic drugs, 696
Parathormone See *Parathyroid hormone*
Parathyroid gland, diseases of, 948-950
during lactation, 996
during pregnancy, 989
physiologic anatomy of, 944
Parathyroid hormone, 877, 943-946
control of calcium ion concentration by, 436, 948
for hypoparathyroidism, 949
in calcium absorption, 793
in phosphate ion concentration, 436
in vitamin D metabolism, 938
Parathyroid poisoning, 949
Paraventricular nuclei, 679
Paravertebral sympathetic chains, 687
Parietal cells, 774
Parietal pathway, 600
Parietal peritoneum, 380
Parkinson's disease, 629-631
Parotid gland, 772
Paroxysmal tachycardia, 201-202
Pars intermedia, 884, 919
Partial pressure(s), 482
alveolar, 485, 490
effect of high, 537-541
high altitude effect of, 528
in deep sea diving, 537-541
Parturition, 991-994
Parvocellular portion, of red nucleus, 636
Passive transport, in renal reabsorption, 401-402
Past pointing, 647
Patent ductus arteriosus, abnormal pressure pulse contours in, 226
at birth, 322, 1000
closure of, 322, 1001
hypertrophy of heart in, 324
machinery murmur in, 323
persistence of, 322-323
surgery for, 323
Pellagra, 868
Pelvic nerves, 687
bladder innervation by, 460, 461
in defecation, 757, 768
Pelvis, renal, 393, 394
Penis, 958-959
Pentagastrin, 777
Pentolinium, 696
Pepsin, 775, 790
Pepsinogen, 774, 775
Peptic cells, 774
Peptic ulcers, 600, 778, 799-801
Peptidases, 785, 872
Peptide chains, 829
Peptide hormones, 878
in gastrointestinal blood flow, 342
Peptide linkage, 32-33, 787, 790, 829
Peptidyl transferase, 32
Periaqueductal gray area, in pain control system, 596-597
Pericardial cavity, 380
Pericardial fluid, 191
Perilymph, 739
Perimetry, 728
Periodic breathing, 514-515
Periodontal membrane, 951
Peripheral nervous system, in alkalosis, 449
Peripheral regulation, of food intake, 864-865
Peripheral resistance See *Vascular resistance, total peripheral*
Peripheral resistance unit, 211
Peripheral venous pressure, 221-222
Peristalsis, 758-759
antral, 763
in colon, 767-768
in esophagus, 761
in small intestine, 766
in stomach, 763
primary vs secondary, 761
Peristaltic constrictor rings, 763
Peristaltic reflex, 759
Peristaltic rush, 766
Peristaltic waves, 766
Peritoneal cavity, 380-381
Peritoneointestinal reflex, 769
Peritonitis, abdominal muscle spasm in, 617

Pentonitis (*Continued*)
chemical, 801
septic shock from, 333
Pentubular capillary network, 394, 395-396
Perivascular space, 375-376
Periventricular nuclei, in pain control system, 597
in punishment, 680, 681
Permeability, capillary, 242
in edema, 369-370
in potential spaces, 380
in shock, 329
relative, 350-351
of membrane, 93
selective, of cell membrane, 95
of protein channels, 91
Permissive level of pumping, 272-274, 276
Peroxidase, 898
Peroxisomes, 16
formation of, 22
in neutrophils and macrophages, 54
pH, 439
calculation of, 440
intracellular, 439
measurement of, 450
of blood, 502
determination of, 517
in respiratory control, 508, 509-510, 511
of extracellular fluids, 442-443
of urine, 448
Phagocytic vesicle, 54
Phagocytosis, 19, 20
by neutrophils and macrophages, 51, 53-54
in complement cascade, 65
Phagosome, 54
Pharyngeoesophageal sphincter, 760
Pharynx, in swallowing, 760-761
Phasic receptors, 576
pH-bicarbonate diagram, 450, 459
Phenoxybenzamine, 696
Phentolamine, 696
Phenylephrine, 696
Phenylthiocarbamide, 746
Pheochromocytoma, 267
Phlorhizin, 794
Phonation, 478
Phonocardiogram, 154, 318, 319
Phosphagen system, 1009-1010, 1011
Phosphate, absorption of, 937, 939, 945-946
excretion of, 937, 945
fetal metabolism of, 998
in ATP, 23
in bone, 941-942, 944-945
in irreversible shock, 331
reabsorption of, 405, 406
Phosphate bonds, high-energy, 23, 29
Phosphate buffer system, 441, 446
Phosphate concentration, in extracellular fluid, 940, 944-946
in rickets, 950
regulation of, 436
Phosphocreatine, 129, 842, 1010
Phosphofructokinase, 815, 924, 928
Phosphogluconate pathway, 816
Phospholipase, 779, 790
Phospholipids, 824-825
absorption of, 795
digestion of, 789-790
formation of, 825
in blood clotting, 79-80
in cell, 12
Phospholipids (*Continued*)
in cell membrane, 13, 826
in chylomicrons, 818
on platelets, 76, 80
uses of, 825
Phosphoric acid, in DNA, 27-28
Phosphorus, 872
Phosphorylase, 811, 924
Phosphorylation, of glucose, 810
oxidative, 813-814
Photopsins, 716
Physiology, categories of, 2
defined, 2
Physostigmine, 139, 696
Pia mater, 375, 377
Piezoelectric effect, of bone deposition, 943
PIF (prolactin inhibitory factor), 886, 995
Pilocarpine, 696
Piloerection, 855
Pineal gland, 966
Pinocytic channel, 349
Pinocytic vesicle, 19, 20, 349
Pinocytosis, by cell, 19-20
in villi, 791
of secretory vesicles, 138
Pitch, 739
Pituitary diabetes, 889
Pituitary factor, unidentified, 914
Pituitary gland, 884-885
anterior, 884
cell types in, 885
in pregnancy, 989
psychosomatic effects transmitted by, 685
hypothalamic control of, 679, 885-887, 963-964
neural stimulation of, 876
posterior, 884, 893-895
Pituitary hormones, 877, 884
Pituitary stalk, 884, 893
pK, 440
Place principle, 739
Placenta, anatomy of, 985-986
blood flow through, 990
estrogen secretion by, 877, 988
function of, 985-987
permeability of merobrane of, 986-987
progesterone secretion by, 877, 988-989
separation and delivery of, 993
Placental lactogen, 989
Placental villi, 985-986
Planning function, 646
Plaque, dental, 952
Plasma, 88, 208, 381
calcium in, 939-940, 947-948
fluid movement between interstitial fluid and, 352-359
loss of, hypovolemic shock from, 331-332
phosphate in, 940
potassium concentration in, 421
transfusion of, in shock, 334
urea concentration in, 418
volume of, 383, 385
Plasma cells, 51
antibody formation by, 63
genesis of, 52
Plasma clearance, 407-408
Plasma colloid osmotic pressure, 22, 352, 353, 356-357
in edema, 369, 380
in fluid volume excretion, 422-423
Plasma membrane, 121
Plasma proteins, binding of thyroid hormones with, 899
carbon dioxide combined with, 501
cortisol effect on, 915
formation of, 837
in edema, 368-369
in nephrotic syndrome, 457-458
metabolic roles of, 831-834
Plasma substitutes, 334
Plasmablasts, 63
Plasmin, 82-83
Plasminogen, 82
Platelet(s), 51
characteristics of, 76-77
genesis of, 52
in clot retraction, 78-79
life span of, 53
low level of, 83-84
Platelet plug, 76, 77
Plethysmograph, body, 516-517
Pleural cavity, 380
Pleural effusion, 191
Pleural pressure, 467
Pneumonia, 520
Pneumotaxic center, 504, 505
Poiseuille's law, 212
Polar body, 983
Polarography, 517
Polycythemia, 49-50
Polydipsia, 934
Polymenorrhea, 903
Polymorphonuclear cells, 51, 52
Polypeptides, in immunity, 60
Polyphagia, 934
Polyribosomes 32
Polysaccharides, 787
Polyuria 934
Pons, autonomic control by, 695
Pontile nuclei, 640
Pontine reticulospinal tract, 620
Pontocerebellar fibers, 635
Pontocerebellar tracts, 640
P_{O_2}-P_{CO_2}, V_A/Q diagram, 490-491
Pores, capillary, 349, 350-351
cell membrane, 13
nuclear, 18
Portal circulatory system, 340-343
Portal obstruction, 342-343
Portal venous pressure, 342-343
Portal venules, 833
Position sense, 580, 588-589
of scratch reflex, 617
Positive feedback, in brain edema, 340
in control systems, 6
in progressive shock, 328, 330
Positive supportive reaction, 615
Postcentral gyrus, 584, 585, 747
Posterior pituitary gland, 884, 893-895
Posterior pituitary hormones, 877 884
Postganglionic neurons, 687-688
drugs stimulating, 696-697
Postsynaptic inhibition, electrical events, in, 556
receptors in, 550, 552
Postsynaptic membrane, in memory, 659
receptors on, 551-552
synaptic excitation of, 555
Postsynaptic neuron, 549, 551-552
Postsynaptic potentials, excitatory, 555
inhibitory, 556
summation of, 555, 557-558, 559
time course of, 556
Post tetanic facilitation, 540, 658
Post tetanic potentiation theory, of primary memory, 658
Postural reflexes, 615-616

Postural reflexes (*Continued*)
vestibular, 625
Posture, and baroreceptors, 248
cerebellum in, 613-614
Potassium channels, 91
in postsynaptic membrane, 551
in sinoatrial fibers, 166
voltage gated, 106-108
Potassium ion(s), and membrane excitability, 117
excretion of, 420-421
in brain interstitial fluid, 377
in excitatory postsynaptic potential, 553
in heart function, 162
in inhibitory postsynaptic potentials 556
in saliva, 772-773
in vascular control, 243
Nernst potential for, 101, 102, 105
tubular secretion of, 401, 404, 405-406, 434, 911-912
Potassium ion concentration, across neuronal somal membrane, 554
effect on aldosterone secretion 913
regulation of 434-435
Potassium loss, during exercise, 1018
Potassium sodium "leak" channels, 105
Potential energy of pressure, 137
Potential spaces fluid circulation in, 379-381
Power law, 577, 739-740
Power stroke, 125
PP cells, 923, 932
P Q interval, 179
prolonged, 198
P R interval, 179
prolonged, 198
Practolol, 696
Precapillary sphincter, 231, 233, 348
Precentral gyrus, 633
Precipitation, gravitational, 477
of antigen, 61
turbulent, 477
Precursor secretion, of sweat gland, 852
Predictive functions, cerebellum in, 646-647
Prednisone, 910
Preeclampsia 991
Preferential channels, 231, 348
Prefrontal areas, 656
Preganglionic neurons 687-688
Pregnancy, hormonal factors in, 987-990
response of mother's body to, 990-991
toxemia of, 209, 991
Pregnanediol, 973
Preload, 159
Prelumirhodopsin, 714
Prelymphatics, 361
Premature contractions, 175, 199-201
Prematurity, 1004-1005
Premotor area, 632-633, 634
Preoptic area, 679, 850-851
Preprohormone, 878
Prepyriform area, 751
Presbyopia, 701-705
Pressoreceptors, 247
Pressure. See also *Arterial pressure; Blood pressure, Capillary pressure*
atrial, left, 288, 291
right, 221-224, 282-284, 307, 308
barometric, 528
cerebrospinal fluid, 376
[illegible] closing, 214
Pressure (*Continued*)
gas, 482-483, 490
glomerular, 398
interstitial See *Interstitial fluid pressure.*
intra alveolar, 467
intraocular, 378-379
intrapleural, 467
intrarenal, 396
mechanoreceptive sense of, 580
oncotic, 221, 356
osmotic, 95-96, 388
colloid See *Colloid osmotic pressure*
total, 356 389
potential energy of, 157
pulmonary wedge 288
reabsorption, 358
respiratory, 467-468
static, 215
vapor, of water, 482-483
at high altitude, 529
ventricular, 158-159 288
water, 517
wedge, 288
yield, 214
Pressure buffer system 218-219
Pressure difference, effect on diffusion 91
Pressure diuresis 257 258-260
Pressure natriuresis 257 258-260
Pressure pulse(s) in arteries 225-228
in veins, 228-229
Pressure pulse contours 225
abnormal, 226
in pulmonary artery 288
in right ventricle 288
Pressure sodium relationship, 263-264
in essential hypertension, 270
Pressure volume curves See *Volume pressure curves*
Prestitial nucleus 626
Prestriate cortex 727
Presynaptic facilitation in primary memory, 658
Presynaptic inhibition, 556-557
in primary memory, 658
Presynaptic neuron, 549-551
Presynaptic terminals 550-551
acetylcholine secretion by, 137
in memory 659
Pretectal nuclei, 724
Prevertebral sympathetic ganglia, 757
PRF (prolactin releasing factor), 995
Primary auditory cortex, 653, 741-742
Primary ending, 608-609
Primary motor area, 632
excitation of spinal cord by 639-638
removal of, 638
topographical map of, 633-638
Primary olfactory cortex, 751
Primary organizer, of embryo, 38
Primary secretion, of sweat gland, 852
Primary sensory areas, 652-653
Primary somatic sensory area, 652-653
Primary visual cortex, 653, 725-728
Procaine, 117
Procarboxypolypeptidase, 779
Procoagulants, 77
Proerythroblast, 43, 44
Progesterone, 877, 972
chemistry of, 973
degradation of, 973-974
effect on FSH and LH, 977
effect on uterine contractility, 991-992
functions of, 975
in breast development, 974, 994
Progesterone (*Continued*)
in endometrial cycle, 976
produced by ovaries, 968
secreted by corpus luteum, 972
secreted by placenta, 988-989
suppression of fertility by, 981
synthesis of, 973
transport of, 973
Progestins, 972, 973
Proinsulin, 924
Prolactin, 877, 884
in lactation, 994-995
in testosterone production, 963
Prolactin inhibitory factor (PIF), 886, 995
Prolactin releasing factor (PRF), 995
Prolyl hydroxylase, 870
Prometaphase, 37
Promoter 29-30, 33
Pronucleus, 984
Prophase, 36-37
Proportionality factor, 303
Propranolol, 696
for angina pectoris, 303
Proprioceptive sensations, 580, 625
Proprioceptors, in neck, 625
Propriospinal fibers, 607
Propulsive movements, in gastrointestinal tract. See *Peristalsis.*
Propylthiouracil, 905
Prosecretin 780
Prosophenosia, 655-656
Prostaglandins as second messenger, 881
as vasodilator agent, 242-243
fetal, 992
in fever 858
in seminal vesicles, 957
synthesis of, 76
Prostate gland, 954, 957, 964-965
Prostatic ducts, 954
Protamine, 85
Protanope, 718-720
Proteinuria absorption of, 784-790
activator, 34
basic properties of, 829
carbohydrate formation from, 816-817
carrier See *Carrier proteins*
channel, 88-89
complete, 833
composition of, 787
content in different foods, 861
daily requirement for, 862
depletion of, 866, 928-929
digestion of, 790
dynamic action of, 846
effect on metabolic rate, 846
fibrillar, 11
globular, 11-12
histone and nonhistone, 36
in cell, 11-12
in cell membrane, 13-14, 88-89
in interstitial fluid, 363
control of, 365
in edema safety factor, 368
in liver, 915
incomplete, 833
integral, 13-14
obligatory degradation of, 833-834
partial, 833, 862
peripheral, 13, 14
plasma, See *Plasma proteins.*
reabsorption of, 404-405, 406
receptor, 551
in smooth muscle, 144
in synaptic fatigue, 570-571

Protein(s) (*Continued*)
surface, 62–63
repressor, 33–34
storage of, 831
effect of insulin on, 928
structural, 11
transport, 88–89
triglycerides synthesized from, 823
use for energy, 832–833
Protein buffer system, 441–442
Protein channels, bulk flow of water through, 96
diffusion through, 91
gating of, 91–92, 137
in effector organ receptors, 689
in postsynaptic membrane, 551
Protein hormones, 878
Protein kinases, 552
Protein metabolism, 829–834
effect of cortisol on, 915
in Cushing's syndrome, 921
effect of estrogen on, 975
effect of growth hormone on, 887–888
effect of insulin on, 928–929
effect of progesterone on, 975
effect of testosterone on, 834
liver in, 837–838
rate of, 862
Protein sparers, 834, 862
Protein synthesis, effect of estrogens on, 974
effect of growth hormone on, 887–888
effect of insulin on, 928
effect of steroid hormones on, 881
effect of testosterone on, 962
effect of thyroid hormones on, 900
ribosomes in, 21, 31–33
Proteinases, 699
Proteoglycans, in bone, 941
in cell membrane, 14
in interstitial spaces, 370
Proteolytic enzymes, in blood coagulation, 80
in phagocytosis, 54
in saliva, 773
in sperm, 955, 958, 963
Prothrombin, conversion to thrombin, 78
decreased, 83
Prothrombin activator, 78, 79–82
Prothrombin time, 86
Proton acceptor, 438
Proton donor, 438
Protoplasm, 11–12
Protoporphyrin IX, 46
Provitamins, 867
Proximal tubule, 393, 400
absorption through, 402–403, 404–405
hydrogen ion secretion in, 444
osmolal concentration in, 417
osmosis through, 401
potassium reabsorption by, 420
sodium reabsorption by, 419
Pseudopodium, 24–25
Psychosomatic disorders, 684–685
Pteroylglutamic acid. See *Folic acid*.
Ptyalin, 772, 787–788
Puberty, female, 969, 978–979
male, 964
Pudendal nerve, 461, 462
Pulmonary. See also *Lung(s)*; *Respiratory system*.
Pulmonary artery, 287
pressures in, 288, 291
Pulmonary capacities, 470–471
Pulmonary capillaries, 288, 291–294
Pulmonary circulatory reserve, 293
Pulmonary circulatory system, 206
blood flow in, 289–291
obstruction of, 293–294
blood volume of lungs, 288–289
capillary dynamics in, 288, 291–294
in exercise, 290–291
in left heart failure, 290
physiologic anatomy of, 287–288
pressures in, 288
Pulmonary congestion, 309, 312, 471
Pulmonary edema, 292, 372
alveolar fluid in, 293
at high altitudes, 532
death from, 372–373
from capillary damage, 293
in cardiac failure, 309, 312, 320
in valvular disease, 320–321
"interstitial fluid" vs "alveolar," 372
safety factor in, 292–293, 372
Pulmonary embolism, 84, 293
Pulmonary function studies, 471–472
Pulmonary lymphatics, 287–288
Pulmonary membrane, 486–490, 518
Pulmonary reflexes, 250
Pulmonary resistance, at birth, 1001
total, 212
Pulmonary stenosis, 321, 324
Pulmonary valves, 155, 156
Pulmonary ventilation, alveolar, 472–474
artificial respiration, 478–479
at high altitudes, 530
effect of smoking on, 1015
in exercise, 1014
mechanics of, 466–469
minute respiration volume in, 472
pulmonary volumes and capacities in, 470–472
respiratory passageways in, 474–478
Pulmonary volumes, 470, 471
Pulmonary wedge pressure, 288
Pulp, splenic, 55, 343
tooth, 951
Pulsatile output, of heart, 284
Pulse(s), bigeminal, 200
pressure, 225–229
radial, 228, 245–246
weak, 228
Pulse deficit, 228
in atrial fibrillation, 203
in premature atrial contractions, 200
Pulsus alternans, 228
Pulsus paradoxus, 228
Pulvinar, 724
Pump(s), atria as, 153, 154
electrogenic, 103, 104–106
heart as, 153
effectiveness of, 279
lymphatic, 363–364, 365
in drainage of potential spaces, 380
pyloric, 763–764
sodium-potassium, 97–98
in membrane potentials, 103–106, 554–555
ventricles as, 153, 155
venous, 223
Pumping, permissive level of, 272–274, 276
Punishment centers, 680–681
Pupillary aperture, 705
autonomic control of, 732–733
in light and dark adaptation, 717
Pupillary light reflex, 733
Purines, enzyme regulation of, 35
in DNA, 27, 28
Purkinje cell, 641–642
Purkinje fibers, 165, 168
Purkinje system, 167–168
blocks in, 192
in preventing ventricular fibrillation, 174
in synchronous contraction of ventricles, 170
Purpura, thrombocytopenic, 84
Pursuit movements, 730–731
Pus, formation of, 57
Putamen, 626–629
Pyelography, intravenous, 459
Pyelonephritis, 453
Pyloric glands, 774, 775–776
Pyloric pump, 763–764
Pyloroplasty, 801
Pyramidal cells, giant, 635
in cerebral cortex, 651
Pyramidal neurons, dynamic and static, 637
Pyramidal tract, 632, 635, 636
Pyridoxal phosphate, 869
Pyridoxine, 869–870
Pyriform area, 751
Pyrimidines, enzyme regulation of, 35
in DNA, 27, 28
Pyrogens, 858
Pyroglutamyl-histidyl-proline-amide, 903
Pyruvic acid, in alanine formation, 832
in anaerobic glycolysis, 816–818
in energy release, 22, 811–812

Q waves, 154, 177, 186–187
QRS complex, 154, 177
and monophasic action potential, 178
bizarre, 192
current of injury effect on, 192–193
decreased voltage of, 191
in recovered myocardial infarction, 195
increased voltage of, 191
prolonged, 191–192
vectors during, 185–187
QRS vector, mean, 188–191
Q-T interval, 179

R waves, 154, 177, 187
Radial pulse, 228, 245–246
Radiation, and cancer, 38
and thrombocytopenia, 84
at high altitudes, 535
heat loss by, 850–851
in atomic submarines, 543
in space, 535
Radioimmunoassay, 882
Rage, 681
Raising movements, 626
Ramp signal, 505, 506
Raphe nuclei, in depression, 683
in pain control system, 597
in reticular activating system, 668
sleep production by, 672
Rate control, of enzyme-catalyzed reactions, 843–844
Rate receptors, 576
Rathke's pouch, 884
Raynaud's disease, 345–346
Reabsorption, 395, 400
in different tubule segments, in fluid volume regulation,

Reabsorption (*Continued*)
mechanisms of, 400-402
of bicarbonate ion, 406, 445
of potassium, 420, 421
of sodium, aldosterone in, 911-912
by active transport, 400-401
in cortical collecting tubule, 403
in late distal tubule, 403
in loop of Henle, 419
in proximal tubules, 419
of urea, by collecting duct, 404, 418-419
passive mechanism of, 401-402
percentage absorbed, 403
Reabsorption pressure, 358
Reagins, 68
Rebound, in cerebellar disease, 617
of reflexes, 615
Receptive field, 562
Receptive relaxation, 759, 761
Receptor(s), acetylcholine, 689-690
adrenergic, 690
alpha, 298, 337, 690
beta, 298, 337, 690
cold, 603-605, 854
down regulation of, 879
electromagnetic, 572
excitatory, 550, 552
in smooth muscle, 144, 145
Golgi tendon, 588, 589
hormone, 876, 879-880
in pinocytosis, 19
inhibitory, 550, 552
in smooth muscle, 144, 145
irritant, 475, 513
J, 513-514
joint, 588
low pressure, 250
movement, 576
muscarinic, 689-690
muscle, 607-613
nicotinic, 689, 690
of effector organs, 689-690
osmosodium, 429-431
pain, 593-594
phasic, 576
position sense, 588
postsynaptic, 551-552
rate, 576
sensory. *See Sensory receptors*
spindle, 609
stretch, in lungs, 505
tactile, 580-581
thermal, 572, 603-605
tonic, 576
up regulation of, 879
volume, 426-427
Receptor potentials, 573-575
of hair cells, 739
of olfactory cells, 749
of rods, 714-715
of taste cells, 747
Receptor proteins, 551
in smooth muscle, 144
in synaptic fatigue, 570-571
surface, 62-63
Reciprocal inhibition, 615
Reciprocal inhibition circuit, 565, 614
Reciprocal innervation, 615
Recurrent inhibition, 607
Red blood cells, concentration of, 42
destruction of, 48
effect of testosterone on, 962
functions of, 42
hemoglobin concentration in, 42
hemoglobin formation, 46
metabolism in, 46-48
Red blood cells (*Continued*)
mature, 43
production of, 42-45
shape and size of, 42
Red flare, 68
Red nucleus, 636-638
Red pulp, cords of, 55
Red green color blindness, 718-720
"Redout," of vision, 633
Red sensitive pigment, 716
Re entry, 172-175
Reflexes, abdominal compression, 252
arterial pressure, 241
autonomic, 694
Bainbridge, 250
baroreceptor. *See Baroreceptor reflexes*
chemoreceptor, 253
chewing, 759
cord. *See Spinal cord reflexes*
cord righting, 616
cough, 475
fatigue of, 615
galloping, 616
gastrointestinal. *See Gastrointestinal reflexes*
Hering Breuer inflation, 505
load, 610
mark time, 616
mass, 617
micturition, 461-462
myotatic, 609-610, 612, 644
neck, 625
pupillary light, 733
rebound of, 615
sexual, 691, 959
skin temperature, 857
sneeze, 475
spinal. *See Spinal cord reflexes*
stretch, 609-610, 612, 611
tendon, 612-613
ureterorenal, 461
vestibular postural, 625
visceral, 656
volume, 250
withdrawal, 613-614
Reflux, prevention of, 762
Refraction, errors of, 705-706
of light, 700
Refractive index, 700
Refractive power, 702-703
Refractory period, in neuronal excitation, 117
of A-V node, 168
of cardiac muscle, 152
Regression, of tissues, 20
Regulatory mechanisms, 4-5
Regulon, 34
Regurgitation, 318
aortic, 226, 319-320, 324
mitral, 319, 320-321, 324
Rehearsal, 660
Reinforcement, 681
Reissner's membrane, 735
Relaxin, 989-990
Release site, 551
Releasing hormones, 679, 885, 886
REM sleep, 671, 672
Renal. *See also Kidney(s)*
Renal blood flow, 395-396
calculation of, 408
control of, 235, 413
effect on glomerular filtration rate, 399
Renal blood vessels, hemorrhage of, 271
Renal calyces, 394
Renal clearance, osmolar, 417
Renal clearance (*Continued*)
plasma, 407-408
tests of, 458-459
Renal disease, 452-458
hypertensive, 457
vascular, 453-454
Renal failure, 452-457
Renal fraction, 397
Renal function, in cardiac failure, 306-309
in newborns, 1002
in shock, 334
tests of, 458-459
Renal glycosuria, 934
Renal output curve, 258, 259
in hypertension, 267-268
sodium pressure relationship in, 264
Renal pelvis, 393, 394
Renal poisons, 453
Renal pressures, 396
"Renal rickets," 950
Renal tubules. *See Tubules*
Renal-body fluid system, for arterial pressure control, 257-262
in vasoconstrictor hypertension, 267-268
Renin, effect on aldosterone secretion, 913-914
general effect of, 876
in arterial pressure regulation, 242, 254-255, 262-264
in cardiac failure, 306
in glomerular filtration rate, 412-413
in hypertension, 267, 268, 269, 457
Renin substrate, 242, 254
Renointestinal reflex, 760
Renshaw cells, 607
Replacement therapy, in shock, 334
Replication, chromosomal, 36-37
DNA, 35-36
Repolarization, 106
atrial, 187-188
propagation of, 111
ventricular, 187
Repolarization waves, 177-178
Repressor operator, 33-34
Repressor protein, 33-34
Repressor substance, 34
Reproduction, cell, 35-37
in homeostasis, 5
sexual. *See Sexual reproduction*
Reserpine, 696
Reservoirs. *See Blood reservoirs*
Residual body, 20
Residual volume, 470, 471
in asthma, 521
measurement of, 517
Resistance. *See Vascular resistance*
Resistance training, 1013
Resolution, in sexual act, 981
Resonance, 478
Respiration. *See also Breathing*
artificial, 478-479
during pregnancy, 991
effect of thyroid hormone on, 902
effect on swallowing, 761
energy required for, 469
in exercise, 1014-1015
Kussmaul, 934
Respirator, tank, 479
Respiratory acid, 448
Respiratory center, 504-506
chemical control of, 506-508
depression of, 514, 530
excitation of, 5
by hydrogen ion concentration, 439,

Respiratory center (*Continued*) 443-444
feedback gain in, 515
Respiratory control, abnormalities of, 514-515
body temperature in, 513
by panting, 513
chemical, 506-508
during exercise, 511-513
irritant receptors in, 513
"J" receptors in, 513-514
oxygen in, 508-511
respiratory center in, 504-506
vasomotor center in, 513
voluntary, 513
Respiratory distress syndrome, 467, 521, 999-1000
Respiratory exchange ratio, 502, 862
Respiratory insufficiency, absorption of entrapped air in, 525
in asthma, 68, 521
in atelectasis, 294, 520-521
in emphysema See *Emphysema*
in hypercapnia, 523
in hypoxia. See *Hypoxia*
in pneumonia, 520
in tuberculosis, 521-522
methods for studying, 516-517
physiologic types of, 517-518
Respiratory membrane, 486-490, 618
Respiratory muscles, 466-467
paralysis of, 471
Respiratory passageways, functions of, 474-478, 484
Respiratory pressures, 467-468
Respiratory quotient, 862-863
Respiratory rate, 472, 474
pneumotaxic center control of, 505
Respiratory reserve, in patent ductus arteriosus, 323
Respiratory system. See also *Lungs*), *Pulmonary*,
at birth, 998-1000
buffering power of, 444
compensation for acidosis or alkalosis by, 449
fetal, 998
in newborn, 1001
in premature infant, 1004
nutrients supplied by, 4
Respiratory unit, 486
Respiratory waves, 252
Respirometer, 284
Resting expiratory level, 471
Resting potential, 104-106
in cardiac muscle, 151
in skeletal muscle, 127
of neuronal soma, 553-555
of smoothal fiber, 166
Resuscitator, 479
Reticular activating system, 665-668, 741
Reticular formation, 619-620
antigravity function of, 620-621
in motor control, 626
of stereotyped movements, 625-626
in pain sensation, 596
neuronal systems associated with, 667-668
Reticular lamina, 738-739
Reticular nuclei, 590
Reticular substance, in vasoconstriction, 239, 241
Reticulocerebellar fibers, 640
Reticulocerebellar tracts, 635
Reticulocyte, 43
Reticuloendothelial cells, of spleen, 343
Reticuloendothelial system, 54-56
Reticulospinal tracts, 620, 635
Reticulum, endoplasmic. See *Endoplasmic reticulum*.
of blood clot, 78
sarcoplasmic, 122
calcium ion release by, 127-129
in muscle action potential, 127
in smooth muscle, 143
Retina, detachment of, 713
formation of image on, 704
fusion of flickering lights by, 717
lateral inhibition in, 721, 722
light and dark adaptation by, 716-717
neural function of, 720-723
pigment sensitivity of, 716
structural elements of, 711-713
Retinal, 714
Retinitis pigmentosa, 728
Retinol, 867
Retinoscope, 709
Retroperitoneal fibroplasia, 236, 1004-1005
Retropulsion, 763
Re uptake, of neurotransmitter, 552
Reverberatory circuit, 567-569
in primary memory, 658
Reverse transcriptase, 38
Reward centers 680, 681
Reynolds' number, 210
Rh blood types, 72-73
Rh factor, 72
Rheobase, 117
Rheumatic fever, 67, 318
Rhodopsin, 712, 713-715
Rhythim method, 981
Rhythmicity, of excitable tissues, 113-114
of heart, autonomic control of, 170-171
of smoothal fibers, 165-167
Rib cage, in breathing, 466-467
Riboflavin, 868-869
Ribonuclease, 778-779
Ribonucleic acid (RNA), 18, 29
in transcription, 29-31
in translation, 31-33
messenger, 30, 31-32
ribosomal, 30, 31
synthesis of, 29-30
transfer, 30, 31
Ribose, in ATP, 23
in RNA, 29
Ribosomal RNA, 30, 31
Ribisomes, attachment to endoplasmic reticulum, 15, 32
formation of, 18, 31
in cytoplasm, 16
in protein synthesis, 21, 31-33
structure of, 31
Rickets, 946, 950
Right atrial pressure, 221-224
in cardiac failure, 307, 308
in cardiac output and venous return curves, 282-284
Rigidity, decerebrate, 630
in Parkinson's disease, 630
Rigor mortis, 134
RNA. See *Ribonucleic acid (RNA)*.
RNA codons, 29, 30
RNA polymerase, 29-30
RNA splicing, 30
Rods, 711-712
excitation of, 714-715, 720-721
Rods of Corti, 738
Rotational movements, 626
Round window, 736, 737
Rubrospinal tract, 635, 636
Ruffini's end organs, 581
in joints, 588, 589
Rugae, 460

S cells, 780
S waves, 154, 177, 187
Saccades, 730
Saccharin, 745
Saccule, 621, 623
Safety factor, at neuromuscular junction, 133
cardiac, 170
edema, 292-293, 368, 372
in impulse propagation, 111, 117
Salbutamol, 696
Saline solution, 391
Saliva, secretion of, 772-774
Salivary ducts, 772
Salivary glands, 770, 772
effect of aldosterone on, 912
Salivatory nuclei, 748, 773
Salpingitis, 982
Salt See also *Sodium*
craving for, 433
in exercise, 1017-1018
intake of, control of, 433
during exercise, 1017-1018
in arterial pressure regulation, 259-284
in hypertension, 265
retention of, 262-263
in cardiac failure, 306
in Goldblatt hypertension, 268-269
Salt sensitivity, 270
Salt tablets, 1017
Saltatory conduction, 114-115
Salty taste, 745
Sarcolemma, 121
Sarcomere, 122
Sarcoplasm, 122
Sarcoplasmic reticulum, 122
calcium ion release by, 127-129
in muscle action potential, 127
in smooth muscle, 143
Satiety, 863
Satiety center, 679, 863
Satisfaction, 680
Scala media, 735, 739
Scala tympani, 735-736, 739
Scala vestibuli, 735-736, 739
Scalenl, 466
Schizophrenia, dopamine system in, 684
Scopolamine, 696
Scotomata, 728
Scotopsin, 714
Scratch reflex, 617
SCUBA diving, 542
Scurvy, 870
Second messenger, for hormone mediation, 880-881
in parathyroid stimulation, 946
in smooth muscle contraction, 145
in steroid production, 918
in thyroid hormone secretion, 903
Secondary auditory cortex, 741-742
Secondary ending, 608-609
Secondary sensory areas, 653-654
Secondary visual cortex, 727-728, 730
Secretagogues, 776
Secretin, chemical composition of, 777
in gastrointestinal blood flow, 3
in gastrointestinal motility, inhibition of gastric secretion

Secretin (*Continued*)
local nature of, 876
regulation of bile secretion by, 783
regulation of pancreatic secretion by, 780
Secretion, gastrointestinal See *Gastrointestinal secretion*
glandular, energy for, 841
hormonal, 878-879
rate of, 879, 883
tubular See *Tubular secretion*
Secretory vesicles, 16
for hormones, 878
formation of, 22
in glandular cells, 771
Segmentation contractions, 765, 766
Seizures, 674-675
Self-excitation, 165-166
Semen, 957
Semicircular canals, 621-625
Semilunar valves, 155, 156, 316-317
Seminal vesicles, 954, 957
Seminiferous tubules, 954
Sensation, deep, 580
exteroreceptive, 580, 625
mechanoreceptive See *Mechano receptive sensation*
modalities of, 573
pain See *Pain*
proprioceptive, 580, 625
segmental fields of, 591
sensitivity to, 590-591
sexual, female, 980
male, 959
thermal, 580, 590, 603-605
vestibular, 621-625
vibratory, 580, 582, 588
visceral, 680
Sensory aphasia, 662
Sensory area, of vasomotor center, 239
primary, 652-653
secondary, 653-654
Sensory association areas, 653-654
Sensory engrams, of motor movements, 648
Sensory feedback control, of motor functions, 637, 648-649
Sensory receptors, 546-547
adaptation of, 575-576
differential sensitivity of, 572-573
joint, 588
movement, 576
pain, 593-594
phasic, 576
position sense, 588
rate, 576
tactile, 580-581
thermal, 603-605
tonic, 576
types of, 572-573
Sensory stimuli, strength of, psychic interpretation of, 576-577
transduction into nerve impulses, 573-576
Sensory terminal, 659
Septal defect, 321, 323
Septic shock, 326, 333
Septum, 676
Series elastic component, of muscle contraction, 130-31
Serosa, 754
Serotonin, as neurotransmitter, 553
as vasodilator agent, 242
in depression, 683-684
in neostriatum-substantia nigra [illegible], 627
Serotonin (*Continued*)
in pain control system, 597
in sleep production, 672
Serous membrane, 380
Sertoli cells, 954-956, 964
Serum, in clot formation, 78-79
Serum cholinesterase, 689
Servo control of muscle contraction, 611
cerebellum in, 648
of muscle tension, 613
Servo assist mechanism, 611
Set point, for temperature control, 856-857
Seventh nerve, 687
Sex chromosomes, 955, 983
Sex hormones, female, 968-969
in athletic performance, 1008-1009, 1018
male, 960-961
Sexual act, female, 980-981
male, 958-960
Sexual characteristics, female, 974-975
male, 961-963
Sexual cycle, female. See *Ovarian cycle*
Sexual drive, hypothalamus in, 680
neuronal stimulus for, 959
sex hormones in, 684
thyroid hormone in, 903
Sexual functions, female See *Ovarian cycle*
male, abnormalities of, 964-966
control of, 963-964
with age, 964
Sexual organs, female, 968
effect of estrogens on, 974
male, 951
Sexual reflexes, 694, 959
Sexual reproduction, fertilization of ovum in, 983-985
hormones in, female, 968-969, 972-979, 987-990
male, 960-964
ovarian cycle in, 969-972
placenta in, 985-987
sexual act in, female, 980-981
male, 958-960
spermatogenesis in, 954-958
Sexual sensation, female, 980
male, 959
SH (somatotropic hormone) See *Growth hormone (GH)*
Shingles, 601
Shivering, 855, 857
Shock, acidosis in, 330
anaphylactic, 332-333, 335
blood transfusion in, 334
cardiac, 278, 300
cardiogenic, 300, 310-311
circulatory. See *Circulatory shock*
compensated, 326-328
effects on body, 333-334
endotoxin, 333
hemorrhagic, 327-332, 333-335
histamine, 333
hypoglycemic, 926
hypovolemic, 327-332, 333, 335
insulin, 936
irreversible, 327, 331
neurogenic, 332, 335
nonprogressive, 326-328
progressive, 327-330
renal, 334
septic, 326, 333
spinal, 618
traumatic, 335
treatment of, 334-335
Shock lung syndrome, 330
Shock therapy, 684
Shunts, arteriovenous, cardiac failure from, 310, 314-315
in cardiac output, 275, 278
left-to right, 321, 322-323
physiologic, 491, 492, 518
hypoxia caused by, 524
in emphysema, 520
right to left, 321-322, 323-324
Sialic acid, synthesis of, 21
Sickle cell anemia, 48-49
Signal(s), 562
amplification of, 564-565
continuous, 569
convergence of, 565
divergence of, 561-565
prolongation of, 567-569
ramp, 505, 506
rhythmic, 569
strength of, 562-563
transmission of, in nerve tracts, 562-563
in nerve trunks, 114-115
in neuronal pools, 563-569
Silicone, as anticoagulant, 85
Simple cells, 726
Sinoatrial block, 198
Sinoatrial fibers, rhythmicity of, 165-167
Sinoatrial node, 165-167
as pacemaker, 169-170
Sinus arrhythmia, 197-198
Skeletal muscle, abnormal functioning of, 133-135
after denervation, 134
anatomy of, 120-122
atrophy of, 134
blood flow through, 336-338
contraction of, 131-133
fasciculation of, 135
fatigue of, 133
fibrillation of, 135
hypertrophy of, 133-134
in arterial pressure control, 252
in familial periodic paralysis, 134
in rigor mortis, 134
length of, 126, 133
muscle tone in, 132-133
Skeletal nerves, in arterial pressure control, 252
psychosomatic disorders and, 684-685
sympathetic nerve fibers in, 687
Skin, circulation in, 343-345, 850
color vs temperature of, 345
estrogen effect on, 975
glabrous, 580
histiocytes in, 55
temperature receptors in, 854
testosterone effect on, 962
Skin temperature, and blood flow, 344-345
and skin color, 345
in temperature regulation, 856-857
Skin temperature reflexes, 857
Sleep, 670
and wakefulness cycles, 673
basic theories of, 672-673
desynchronized, 672
effect of thyroid hormone on, 902
effect on metabolic rate, 846-847
electroencephalographic changes in, 671
hypothalamus in, 681
neuronal mechanisms of, 672-673
paradoxical, 672
physiological effects of, 673-674

Sleep (*Continued*)
REM, 671, 672, 673
slow wave, 671–672
Sleep spindles, 671
Sliding filament mechanism, 122-123
Slit-pores, 397, 398
Slow channels, in action potential, 110, 113
in cardiac muscle, 151
in sinoatrial node rhythmicity, 166
Slow reactive substance of anaphylaxis, 475, 521
Slow wave potentials, 142–143
Slow wave sleep, 671–672
Slow waves, of gastrointestinal smooth mucle, 755, 762–763, 765
Small intestine, absorption in, 791–796
digestion in, of carbohydrates, 788
of fats, 788–789
of *proteins*, 790
disorders of, 801–802
inhibition of gastric secretion by, 778
movements of, 765–767
obstruction of, 804–805
secretions of, 784–785
Smell, 748–751
Smoking, and exercise, 1015
atherosclerosis with, 827
chronic emphysema from, 518–520
effect on ventilation perfusion ratio, 491–492
Smooth muscle, action potentials in, 140, 141–143
contraction of, 140–141
tonus, 146
without action potentials, 145–146
excitation-contraction coupling in, 143–144
in gastrointestinal tract, 754–756, 768
mechanical characteristics of, 148
membrane potentials in, 141
multiunit, 139–140
muscle tone of, 148
neuromuscular junctions of, 144–145
shortening of, 148
single-unit, 140
size of, 139
stress-relaxation of, 146
types of, 139–140
unitary, 140
visceral, 140
Sneeze reflex, 475
Sodium appetite, 433
Sodium bicarbonate, for acidosis, 450
Sodium channels, fast, 110, 113, 151, 166
in action potential, 106, 107, 109–110, 113
in cardiac muscle, 151–152, 166
in postsynaptic membrane, 551
in sinoatrial node rhythmicity, 166
in smooth muscle, 142
gastrointestinal, 756
inactivation of, 107
selective permeability of, 91
voltage-gated, 106, 107, 109–110
Sodium cotransport, 99
in gastrointestinal tract, 794, 795
in renal tubules, 401
Sodium countertransport, 99
in renal tubules, 444
Sodium gluconate, for acidosis, 450
Sodium ion(s). See *Salt*.
absorption of, 792–793, 796
excretion of, 419–420
in excitatory postsynaptic potential, 555
Sodium (*Continued*)
in heart function, 162
in saliva, 772–773
in vascular control, 243
Nernst potential for, 102, 105
tubular reabsorption of, 400–401, 403, 405, 419, 911–912
Sodium ion concentration, across neuronal somal membrane, 554
aldosterone secretion and, 911, 914
in extracellular fluid, 428–433, 894
Sodium lactate, for acidosis, 450
Sodium-potassium pump, 97–98
in membrane potential, 103
in resting membrane potential, 104–106
of neuronal soma, 554–555
Sodium-pressure relationship, 263–264
in essential hypertension, 270
Solation, 24
Solubility coefficient, 482
Solutes, excretion of excess, 414–417
Soma, 550
during and following action potential, 559
resting membrane potential of, 553–555
Somatic association areas, 586–587
Somatic sensation(s) See *Mechanoreceptive sensation, Pain Thermal sensation*
Somatic senses, classification of, 580
Somatic sensory areas, primary, 652–653
secondary, 653–654
I and II, 584–585, 586
Somatic sensory association area, 586–587, 653–654
Somatic sensory cortex, 584–587, 652–654
Somatointestinal reflex, 769
Somatomedins, 889
Somatostatin, in glucagon secretion, 932
in growth hormone secretion, 890
in inhibition of gastric secretion, 778
in insulin secretion, 932
in TSH secretion, 904
Somatotropes, 885
Somatotropic hormone (SH) See *Growth hormone (GH)*
Somatotropin See *Growth hormone (GH)*
Somites, 38
Sound, attenuation of, 735
conduction to cochlea, 734–735
direction of, 742–743
frequency of, 737–740, 742
hearing threshold for, 740
loudness of, 739–740
patterns of, 742
transmission of, in cochlea, 737–738
through bone, 735
Sour taste, 745
Space of Disse, 835–836
Space physiology, acceleratory forces in, 532–534
"artificial climate" in, 535
radiation hazards in, 535
temperature problems in, 534–535
weightlessness in, 536
Spatial orientation, in dorsal column-lemniscal system, 584
in somatic sensory area I, 585
transmission of, 565–567
Spatial summation, in spinal transmission, 562–563
of auditory vibrations, 739
Spatial summation (*Continued*)
of postsynaptic potentials, 557–558
of thermal sensations, 601–605
Speech, Broca's area in, 634
respiratory system in, 477–478
Sperm, abnormal, 958
flagellum of, 25
formation of, 955
maturation of, 956
physiology of, 956–957
storage of, 956
Sperm count, 958
Spermatids, 954
Spermatocytes, 954
Spermatogenesis, 954–958
feedback control of, 964
Spermatogonia, 954
Spermiation, 956
Spherocytosis, hereditary, 48
Sphincter of Oddi, 782
Sphingomyelin, 114, 824, 825
Spike and dome pattern, 675
Spike potential, 112
in cardiac muscle, 151
in gastrointestinal muscle, 755–756
in smooth muscle, 142
Spinal animal, 608
Spinal cord, body temperature receptors in, 854
functions controlled by, 548
in motor pathways, 620, 636–638
in pain, 594–595, 597
multisegmental connections in, 607
transection of, 617–618
Spinal cord injuries, paralysis of defecation in, 803
Spinal cord reflexes, autonomic reflexes, 616
crossed extensor reflex, 614–616
effect of cord transection on, 617–618
fatigue of, 615
flexor reflex, 613–614
locomotor reflexes, 615–618
muscle receptors in, 607–613
muscle spasm produced by, 616
organization for, 606–607
postural reflexes, 615–618
reciprocal inhibition and reciprocal innervation, 615
scratch reflex, 616
sexual, 959
Spinal motor neurons, stimulation of, 637–638
Spinal nerves, in circulatory regulation, 237
sympathetic fibers in, 687
Spinal shock, 618
Spindle, mitotic, 16, 36
Spindle receptor, 609
Spinocerebellar tracts, 640–641
Spinocervical pathway, 584, 589
Spinoolivary pathway, 641
Spinoreticular pathway, 590, 641
Spinotectal tract, 590
Spinothalamic tracts, 590
Spiral ganglion of Corti, 738, 740
Spirometry, 470–472, 516–517
Spironolactone, 460
Splanchnic circulation, 340–343
Spleen, as hemopoietic organ, 343
as reservoir, 343
blood-cleansing function of, 343
circulation in, 343
macrophages in, 55–56
red blood cell fragmentation in reticuloendothelial cells of,

Splenectomy, for thrombocytopenia, 81
Splenic pulp, 55, 341
Sports physiology: body fluids and salt in exercise, 1017-1018
  body heat in exercise, 1017
  cardiovascular system in exercise, 1015-1017
  drugs and athletes, 1018
  female and male athletes, 1008-1009
  muscles in exercise, 1009-1014
  respiration in exercise, 1014-1015
Sprue, 801-802
Squeeze, in deep sea diving, 541-542
Squint, 731
S-T segment, 193
Staircase effect, 132
Stapedius muscle, 735
Stapes, 734, 736, 737
Starches, 787-788
Starling equilibrium, for capillary exchange, 358-359
Starvation, 866
  effect on protein degradation, 833-834
  ketosis in, 822
Statoconia, 621
Stearic acid, 818
Steatorrhea, 802
Stellate cells, 643, 651
Stem cells, 41, 61
Stenosis, 318
  aortic, 319-320, 321, 321
  congenital, 321, 324
  mitral, 319, 320-321, 321
  pulmonary, 321, 324
Stepping movements, 616
Steroidin, 796, 839
Steroxilia, 622
Stereopsis, 708, 725, 731
Sterility, female, 981-982
  male, 957-958
Sternocleidomastoid muscles, 466
Steroid hormones, chemistry of, 879, 910
  in protein synthesis, 881
Stigma, 970-971
Stimulatory field, 563
Stimulus, excitatory, 564
  strength of, and receptor potential, 574
    psychic interpretation of, 576-577
  subthreshold, 564
  threshold, 564
  transduction of, 573-576
Stimulus artifact, 118
Stokes-Adams syndrome, 170, 180
Stomach. See also Gastric
  anatomy of, 762
  basic electrical rhythm of, 762-763
  digestion in, of carbohydrates, 788
    of proteins, 790
  disorders of, 798-801
  emptying of, 763-765
  motor functions of, 762-765
  secretions of, 774-778
  storage function of, 762
Strabismus, 731
Streptokinase, 83
Stress, bone, 913
  cortisol in, 918
  effect on ACTH secretion, 918
Stress response, 694, 695
Stress relaxation, of blood vessels, 216, 255, 284
  of smooth muscle, 146
  of tissue spaces, 365
  reverse, 146
    in shock, 328
  [illegible] in smooth muscle excitation, 143
    [illegible], 209
Stretch receptors, in lungs, 505
Stretch reflex, 609-610, 612
  cerebellum in, 614
Stria vascularis, 739
Striate area, 727
Striate body, 627
"Stroke," 638
Stroke volume output, 155, 157
  and pressure pulse, 225-226
  in exercise, 1016
  in valvular heart disease, 319
Stroke work output, 157
Stroma, ovarian, 968
Strychnine, 561
Subarachnoid space, 375
Subcallosal gyrus, 676, 683
Subcutaneous venous plexus, 344, 345
Subendocardial arteries, 296
Subendocardial muscle, 296
  myocardial infarction in, 300
Subliminal zone, 564
Sublingual gland, 772
Submandibular gland, 772
Submarines, 542, 543
Submucosa, 754
Submucosal plexus, 756-757
Subnuclear clefts, 137
Substance P, as hormone, 886
  as neurotransmitter, 553
  in neostriatum-substantia nigra pathway, 627
Substantia gelatinosa, 595
Substantia nigra, 626
  damage to, 629-630
  in reticular activating system, 667
  inhibitory pathway between neostriatum and, 627
Substrate, in metabolic reactions, 814
Subthalamus, 626, 629
Subthreshold stimulus, 564
Subthreshold zone, 564
Sucrase, 785, 788
Sucrose, as diuretic, 459
  digestion of, 787
  in taste sensation, 745
Sulfates, reabsorption of, 405, 406
Summation, asynchronous, 132
  in signal transmission, 562-563
  multiple motor unit, 131-132
  of auditory vibrations, 739
  of individual contractile pulses, 146
  of muscle contraction, 131-132
  of postsynaptic potentials, 555, 557-558
  of thermal sensations, 604-605
  spatial, 131
  temporal, 131
  wave, 131, 132
Superior cervical ganglion, 732
Superior colliculus, 724, 730
Superior oblique, 728-729
Superior olivary nucleus, 740, 743
Superior rectus, 728-729
Superoxide, in neutrophils and macrophages, 54
Supplemental motor area, 633-634
Suprachiasmatic nucleus, 724
Supraoptic nuclei, 679, 893
Supraventricular tachycardia, 202
Surfactant, in alveoli, 467-468, 521
  in hyaline membrane disease, 1001
  lack of, in atelectasis, 520-521
Surround inhibition, 588
Sustentacular cells, 746, 749
Swallowing, 760-761
  disorders of, 798
Sweat glands, 852
Sweat glands (*Continued*)
  effect of aldosterone on, 912
  sympathetic stimulation of, 691
Sweating, autonomic regulation of, 852-853
  during exercise, 1017-1018
  in regulation of body temperature, 854, 856-857
  local, 857
  vasodilator mechanism of, 345
Sweet taste, 745
Sympathectomy, for Raynaud's disease, 346
Sympathetic blocking agents, for angina pectoris, 303
Sympathetic ganglia, 687
Sympathetic nervous system, adrenal medullae in, 692-693
  alarm or stress response of, 694, 695
  cholinergic and adrenergic fibers in, 688-689
  circulatory regulation by, 237-240, 25[illegible]
    cerebral, 340
    coronary, 298
    during exercise, 337-338
    gastrointestinal, 342
    in liver, 341
    in skin, 344-345
    pulmonary, 289
    skeletal muscle, 337-338
  compensation by, for acute cardiac failure, 305-306, 307, 313
    for shock, 327, 328
  control of heart by, 160-61, 170-171
  denervation supersensitivity with, 693-694
  effect on metabolic rate, 846
  effects on specific organs, 690-692
  excitatory and inhibitory effects of, 69[illegible]
  gastrointestinal innervation by, 757
  in cardiac output, 273, 276-277, 282-283
  in control of bronchioles, 475
  in emission and ejaculation, 959-960
  in exercise, 276-277, 337-338
  in fluid volume excretion rate, 423
  in gastrointestinal secretion, 771
  in glomerular filtration rate, 400
  in hypertension, 269
  in volume-pressure relationships, of arterial and venous systems, 215
    of entire circulation, 216
  mass stimulation by, 694
  physiologic anatomy of, 686-687
  receptors of effector organs in, 689-69[illegible]
  stimulus rate in, 693
  temperature regulation by, 241
Sympathetic tone, 693
Sympathetic transmitters, 688
Sympathetic vasoconstrictor system, 238-240
  in arterial pressure control, 253-255
  in skeletal muscle, 337
  in skin, 344
Sympathetic vasodilator system, 240
  during exercise, 277
  in skeletal muscle, 337
  in skin, 345
Sympatholytic drugs, in shock, 334
Sympathomimetic drugs, 690, 696
  in shock, 334, 335
Synapse(s), chemical, 549-550
  defined, 548
  electrical, 549
  facilitation of, 548
    in memory, 658-660
    post tetanic, 560, 658

Synapse(s) (*Continued*)
inhibitory, 556
neuromuscular junction as, 138
presynaptic, 556-557
physiologic anatomy of, 550-552
Synaptic bodies, 720
Synaptic cleft, 137, 550
Synaptic delay, 561, 567
Synaptic fatigue, 560
in duration of reverberation, 568
in nervous system stabilization, 570-571
Synaptic gutter, 137
Synaptic knobs See *Terminal(s)*
Synaptic modulators, 552
Synaptic transmission, acidosis effect on, 561
alkalosis effect on, 560-561
drug effect on, 561
firing rate in, 559-560
hypoxia effect on, 561
information processing role of, 547-548
neuronal excitation in, 553-556
neuronal inhibition in, 556-557
role of dendrites in, 558-559
special characteristics of, 560-561
summation of postsynaptic potentials in, 557-558
transmitter substances in, 552-553
Synaptic trough, 137
Synaptic vesicles, 550
recycling of, 551
Syncope, carotid sinus, 249
vasovagal, 241, 332
Syncytium, cardiac muscle as, 150-151
gastrointestinal smooth muscle as, 755
of trophoblasts, 985, 988
ventricular, 151
visceral smooth muscle as, 140
Synovial cavities, 381
Synthesis, cellular, 21-22
protein, 21, 31-33
RNA, 29-30
Systemic circulation, 206, 218-219
Systole, 153
cardiac muscle compression in, 295
in clinical usage, 157
ventricles in, 155
Systolic stretch, 300

T-1824, 385
T cell(s), 54
antigen receptors on, 66
attributes of, 65-66
cell activation by, 63
clones of, 62-63
cytotoxic, 66
helper, 63, 66
origin of, 61
preprocessing of, 61-62
regulatory, 66, 67
specificity of, 62
suppressor, 66, 67
T cell marker, 62-63
T tubules (transverse tubules), in cardiac muscle contraction, 152-153
in skeletal muscle contraction, 127
T vectorcardiogram, 188
T waves, 154, 177
abnormalities in, 195-196
and monophasic action potential, 178
atrial, 178
electrocardiogram of, 187-188
ventricular, 154, 178

Tabes dorsalis, 462
Tachycardia, 175
electrocardiogram in, 197
paroxysmal, 201-202
supraventricular, 202
ventricular, 175
Tachypnea, 516
Tactile receptors, 580-581
Tactile sensations, detection and transmission of, 580-581
Target tissues, 876
Taste, 745-748
Taste blindness, 746
Taste buds, 745, 746-747
Taste reflexes, 748
Tectorial membrane, 738-739
Tectospinal tract, 620
Teeth, physiology of, 950-953
Telophase, 37
Temperature See also *Body temperature*
core, 849, 854
effect on heart, 162
effect on spermatogenesis, 958
extreme cold, 856, 859-860
in aviation and space physiology, 534-535
in TRH and TSH secretion, 904
sensation of, 590, 603-605
skin, 344-345, 856-857
surface, 849, 854
Temperature regulating centers, 853-854, 858
Temporal cortex, 727
Temporal summation, in signal transmission, 563
of postsynaptic potentials, 558
Tendon reflex, 612-613
Tenase cofactor, 767
Tensor tympani muscle, 734, 735
Tentorial opening, 375
Teratoma, 966
Terminal(s), at neuromuscular junction, 136
facilitator, 659
presynaptic, 550-551
acetylcholine secretion by, 137
in memory, 659
sensory, 659
Terminal cisternae, 127
Terminal knobs See *Terminal(s)*.
Testicles, descent of, 958, 961
development of, 988
hormones of, 877
tumors of, 965-966
underdeveloped, 965
Testosterone, 877
and athletic performance, 1008, 1009, 1018
and secondary sexual characteristics, 961-963
chemistry of, 960
effect on libido, 684
effect on metabolic rate, 846
effect on protein metabolism, 834
in fetal development, 961
in spermatogenesis, 956
mechanism of action of, 963
metabolism of, 960-961
secretion of, 960
feedback control of, 964
regulation by LH, 963
Tetanization, 132
Tetany, from hypocalcemia, 940
in alkalosis, 449
in rickets, 950
Tetracaine, 117

Tetraethylammonium ion, 108, 696
Tetralogy of Fallot, 323-324
Tetrodotoxin, 108
Thalamic association areas, 653
Thalamic nuclei, connection with basal ganglia, 627
in limbic sytem, 676
in Parkinson's disease, 630-631
Thalamic syndrome, 601
Thalamocortical system, 652, 668-669
Thalamus, and cerebral cortex, 652
in pain sensation, 595-596
in somatic sensation, 590
in taste, 747
Thalassemia, 49
Thebesian veins, 295
Theca, 970
Thermal receptors, 572, 603-605
Thermal sensation, 580, 590, 603-605
Thermode, 853
Thermogenesis, chemical, 855, 856
nonshivering, 846
Thermostat, hypothalamic, 853-857, 858
Theta waves, 669
Thiamine, 867-868
Thick segment, of loop of Henle, 394, 403, 417
Thin segment, of loop of Henle, 393, 403
Thinking See *Intellectual functions*
Thiocyanate ions, 773, 904-905
Third nerve, 687, 732
Thirst, 431-433
Thirst center, 431, 879
Thoracic duct, 361
Thorax, expansibility of, 468-469
Thoughts, 656-661
elaboration of, 656
sequencing of, 656
transfer to opposite hemisphere, 662-663
Threshold, for excitation, 116
for hearing, 740
of audibility, for heart sounds, 316
pain, 592-593
Threshold stimulus, 564
Threshold voltage, of unmyelinated fiber, 166
"Thrill," 319
Thrombin, 78, 79-80
Thrombocytes See *Platelet(s)*
Thrombocytopenia, 80, 83-84
Thrombocytopenic purpura, 84
Thromboembolic conditions, 84
Thromboplastin, 79-80, 825
Thrombosis, coronary, 195
femoral, 84
of minute vessels, in shock, 329
Thromboxane A, 77
Thrombus(i), 84, 299
Thymine, in DNA, 27, 28
Thymopoietin, 63
Thymus gland, in tolerance, 67
lymphocyte processing in, 61, 62
Thyroarytenoid muscles 478
Thyroglobulin, 897, 898 899
Thyroid cartilage, 478
Thyroid gland, diseases of, 905-908
during pregnancy, 989
physiologic anatomy of, 897
Thyroid hormones See also *Calcitonin*, *Thyroxine*, *Triiodothyronine*
effect on fat utilization, 824, 825
effect on growth, 901
effect on metabolic rate, 684, 846, 877, 900-901
formation of, 897-899
general effect of, 876
in cell nucleus, 881-882

Uracil, in RNA, 29
Urate ions, reabsorption of, 405, 406
tubular secretion of, 401
Urea, as diuretic, 459
concentration of, antidiuretic hormone effect on, 415
in blood, 459
in tubules, 406
excretion of, 417-419
formation of, 833, 837
in chronic renal failure, 454, 455
reabsorption of, 405
by collecting duct, 404
by loop of Henle, 418-419
passive mechanism of, 401-402
Uremia, 454-455
metabolic acidosis in, 448
Ureter(s), 461
Ureterorenal reflex, 461
Urethra, 460, 954
Urethral glands, 954
Uric acid, in renal failure, 455
Uridine diphosphate glucose, 810
Urinary bladder. See *Bladder*
Urination, 461-462
Urine, acid or alkaline, 439
composition of, 395, 407
concentrated, 414-417
dilute, 413-414
during pregnancy, 991
formation of, 400-407
pH of, 448
renal function tests of, 458-459
specific gravity of, 459
sugar in, 934
transport of excess hydrogen ions into, 448-447
transport through ureters, 481
Urine output, 421-424
in blood volume regulation, 426
in chronic renal failure, 454
measurement of, 459
Urobilin, 796, 838
Urobilinogen, 838-840
Urogenital diaphragm, 460
Urokinase, 82
Uriticaria, 68
Uterine arteries, 985
Uterine contractility, during labor, 992-994
near term, 991-992
regulation by hypothalamus, 679
"Uterine milk," 976, 984
Uterine pain, 601
Uterine veins, 985
Uterus, anatomy of, 968
effect of estrogens on, 974
effect of oxytocin on, 895, 992
implantation of zygote in, 984-985
involution of, 994
progesterone effect on, 975
Utilization coefficient, 497
Utilization time, 117
Uticle, 621, 623

V wave, 154, 155, 228-229
Vaccination, 63, 67
Vagina, 968
Vagotomy, 800-801
Vagus nerve, control of bile release by, 782
control of gastric secretion by, 776
dorsal motor nucleus of, 239
gastrointestinal innervation by, 757
Vagus nerve (*Continued*)
in respiratory control, 505, 508
in swallowing, 761
in taste, 747
stimulation of heart by, 160-161, 170
atrioventricular block from, 198
bradycardia from, 197
in paroxysmal tachycardia, 201
Valvulae conniventes, 791
Valvular heart disease, 318-321, 324
Van Allen radiation belts, 535
van den Bergh reaction, 839-840
Vapor pressure, of water, 482-483
at high altitude, 529
"Varicose veins," 223
Varicosities, 688
Vas deferens, 954
Vasa recta, 394, 414
blood flow in, 396
countercurrent exchange mechanism in, 416
Vascular See also *Blood*
Vascular capacitance See *Vascular compliance*
Vascular compliance, 214-216
and pressure pulse, 225-226
in neurogenic shock, 332
Vascular congestion, pulmonary, 309, 312, 471
systemic, 312
Vascular diameter, 212-213
Vascular disease, of limbs, 345-346
Vascular distensibility, 214-215
Vascular resistance, 208
arteriolar, 219
capillary, 219
conductance and, 212-213
in emphysema, 293
in hepatic vessels, 836
pressure and, 213-214
pulmonary, at birth, 1001
total, 212
systemic, at birth, 1001
in venous return, 281-282
total peripheral, 211-212
arterial pressure in, 246, 260
in cardiac output, 274-275, 278
in hypertension, 265, 266, 267
in venous return, 274-275
units of, 211
venous, 219, 221-222
Vascular spasm, 76
in Raynaud's disease, 345
Vasoactive inhibitory peptide, 886
Vasoactive intestinal peptide, 342, 778
Vasoconstriction, as temperature-increasing mechanism, 855, 857
autonomic control of, 692
by efferent arterioles, 412-413
in cutaneous circulation, 344
local, 857
Vasoconstrictor agents, 241-242
in hypertension, 266-269
Vasoconstrictor area, 238
Vasoconstrictor system, sympathetic, 238-240
in arterial pressure control, 253-255
in skeletal muscle, 337
in skin, 344
Vasoconstrictor tone, sympathetic, 239
Vasodilatation, active, 345
as temperature-decreasing mechanism, 854, 857, 860
autonomic control of, 692
by afferent arterioles, 412, 413
cold induced, 860
Vasodilatation (*Continued*)
in smooth muscle contraction, 145
local blood flow regulation by, 232-233
of coronary arteries, 296-297
of muscles, in exercise, 276
peripheral, 868
venous, cardiac output with, 278
Vasodilator area, 239
Vasodilator drugs, 242-243
for angina pectoris, 302-303
for essential hypertension, 270
Vasodilator substance, 232
Vasodilator system, sympathetic, 240
during exercise, 277
in skeletal muscle, 337
in skin, 345
Vasomotion, 233, 349-350
Vasomotor center, 238-240
depressed, 332
failure, in progressive shock, 328-329
in arterial pressure control, 5, 250-251
in respiratory control, 513
Vasomotor tone, 239
in neurogenic shock, 332
Vasomotor waves, 252-253
Vasopressin. See *Antidiuretic hormone (ADH)*.
Vasovagal syncope, 241, 332
Vector(s), as electrical potentials, 183
defined, 183
direction of, 184
in depolarization, of atria, 187
of ventricles, 185-187
in repolarization, of atria, 188
of ventricles, 187
instantaneous "resultant," 183
projected, 184-185
Vectorcardiogram, 188
Vectorial analysis, of normal electrocardiograms, 185-188
of potentials in different leads, 184-185
of premature ventricular contraction, 201
principles of, 183-185
Vegetative functions, 670
of hypothalamus, 678-679
Vein(s), aqueous, 378
blood reservoir function of, 224-225
cardiac, anterior, 295
damming of, 305
thebesian, 295
central, 835
distensibility of, 214
functions of, 218, 221-225
in nervous control of arterial pressure, 251-252
obstruction of, 278
edema from, 368
pressure pulses in, 228-229
pulmonary, 287
pressures in, 288
sympathetic stimulation of, 338, 342
umbilical, 985
uterine, 985
"Varicose," 223
Velocity, of blood flow, 219
of conduction, 115
in cardiac muscle, 152
in skeletal muscle, 127
of contraction, 126
terminal, 534
Venous admixture, of blood, 494
Venous circulation, cardiac, after acute myocardial i
cross sectional area
volume-pressure c

Venous dilatation, cardiac output with, 278
Venous plexus, subcutaneous, 344, 345
Venous pooling, 332
Venous pressures, 219, 221-224
  peripheral, 221-222
  portal, 342-343
Venous pump, 223
Venous resistance, 219, 221-222
Venous return, defined, 272
  in cardiac failure, 306-307, 310
  in cardiac output regulation, 157-158, 273, 278
  in exercise, 277
  pressure gradient for, 281
  total peripheral resistance in, 217-218
Venous return curves, 279, 280-284
  in cardiac failure, 313-315
Venous sinuses, 55, 343
Venous valves, 223
Ventilation See *Alveolar ventilation, Pulmonary ventilation*
Ventilation perfusion ratio, abnormal, 491-492, 518
  in atelectasis, 521
  in emphysema, 519-520
  in pneumonia, 520
  effect on alveolar gas concentration, 490-492
Ventral respiratory group, 504, 505-506
Ventricle(s), 150
  as pumps 153, 155
  blood flow through, 295-296
  defibrillation of 174-175
  depolarization of 185-187
  dilatation of, 191-192
  emptying of, 155
  filling of, 155 159
  hypertrophy of, 319-320, 324
  QRS complex in, 191-192
  vectorial analysis of, 189-190
  mean electrical axis of, 188-191
  Purkinje system in synchronous contraction of, 170
  repolarization of, 187
  transmission of cardiac impulse in, 168-169
Ventricular contraction, in electrocardiogram 178
Ventricular escape, 170, 199
Ventricular fibrillation, 174-175
  after myocardial infarction, 301
  circulatory arrest from, 335
  electrocardiogram of, 203
Ventricular function curves, 159-160
Ventricular paroxysmal tachycardia, 202
Ventricular pressure, 158-159, 288
Ventricular rhythm, in atrial fibrillation, 174
Ventricular syncytium, 151
Ventricular tachycardia, 175
  paroxysmal, 202
Ventricular volume, 158-159
Ventricular volume curve, 155
Ventrobasal complex, 583, 590
Ventromedial nucleus, behavioral functions of, 679-680
  in food intake, 863
  in growth hormone secretion, 890
Venules, 218
  in microcirculation, 231, 348
  portal, 835
Vermis, 639, 640
  in auditory pathway, 711
Vertebrae, effects of accelerator force on, 533
Vesicointestinal reflex, 769
Vestibular apparatus, 621-623
Vestibular membrane, 735
Vestibular nerve, 622, 623
Vestibular nuclei, 623, 641, 731
Vestibular sensation, 621-625
Vestibulocerebellar fibers, 640
Vestibulospinal tracts, 620
Vibration, sensation of, 580, 582, 588
Vicious circle, 6
Villi, intestinal, 766, 791-792
  placental, 985-986
Virus(es), and cancer, 38
  neutralization of, 65
  vs animal cells, 18-19
Visceral pain, 598-601
Visceral pathway, 599-600
Visceral peritoneum, 380
Visceral reflexes, 686
Visceral sensations, 560
Vision See also *Eye(s)*
  binocular, 708
  color See *Color vision*
  fields of, 728
  fusion of, 725, 731
  in maintenance of equilibrium, 625
  light and dark adaptation in, 716-717
  photochemistry of, 713-717
  "redout" of, 533
Visual acuity, 707-708
  macula in, 711
Visual association areas, 653, 654, 727
Visual cortex, 653, 725-728, 730
Visual images, fusion of, 725, 731
Visual pathway, 724
Visual purple, 714
Visual receptive aphasia, 662
Vital capacity, 471
Vital dyes, 385
Vitamin(s), A, 867
  in retina, 713, 714
  B_1 (thiamine), 867-868
  B_2 (riboflavin), 868-869
  B_6 (pyridoxine), 869-870
  B_{12} (cyanocobalamin), 869
    in fetus, 998
    in red blood cell formation, 45
  C (ascorbic acid), 870
    in fetus, 998-999
    in infant, 1003
  D, 870-871
    deficiency of, 950
    during pregnancy, 990
    effect on bone, 946
    for hypoparathyroidism, 949
    in calcium absorption, 793, 970, 937-939
    in fetus, 999
    in infant, 1003
    in phosphate absorption, 937-939
  daily requirements for, 867
  deficiencies of, 867-871
    in starvation, 866
  E, 871
    in fetus, 999
  in local blood flow control, 234
  in red blood cell formation, 45
  K, 871
    absorption of, 782-783
    deficiency of, 83
    during pregnancy, 990
    formation of, 796
    in fetus, 999
  metabolism of, 867-871
    effect of thyroid hormone on, 901
    fetal, 998-999
  reabsorption of, 404-405
Vitamin(s) (*Continued*)
  storage of, 838, 867
Vitreous humor, 377
Vocal cords, 478
Vocalization, respiratory system in, 477-478
Voice, effect of testosterone on, 962
Voltage clamp, 107-108
Voltage gating, 91, 92
Voltmeter, 103-104
Volume receptors, 426-427
Volume reflex, 250
Volume pressure curves, of arterial & venous circulations, 214-215
  of cardiac cycle, 159
  of entire circulation, 215-216
  of interstitial fluid spaces, 366-367, 368
Voluntary fixation mechanism, 729-730
Voluntary motor activity, initiation of 649
Voluntary muscle control, cerebellum in, 644
Vomiting, 803-804
  acidosis from, 448
  alkalosis from, 449
Vomiting center, 803

Wakefulness, cerebral cortex in, 548
  cycles between sleep and, 673
  electroencephalographic changes in, 871
  hypothalamus in, 681
  reticular activating system in, 665-668
"Walk-along" theory, of contraction, 124-125
Walking movements, 816, 626
Warfarin, 85
Warmth receptors, 603-605, 854
Water, absorption of, 792, 796
  body, intake vs. output of, 382-383
    regulation by hypothalamus, 679
    total, 382, 386
  bulk flow of, 96-97
  diffusion of, net, through cell membrane, 94-97
    through capillary membrane, 350
    through lipid bilayer, 90
  excretion of excess, 413-414
  formed in oxidative phosphorylation, 814
  heat loss by exposure to, 851
  heavy, 386
  in cell, 11
  intake of, in arterial pressure regulation, 259-262
    vs output, 382-383
  retention of See *Fluid retention.*
  secretion in gastrointestinal tract, 771-772, 779-780
  tritiated, 386
  tubular reabsorption of, antidiuretic hormone in, 415-416
    in different segments, 404, 406-407
    passive mechanism of, 401
  vapor pressure of, 482-483
    at high altitudes, 529
Water diuresis, 262, 430-431
Water pressure, 537
Weber Fechner principle, 577
Wedge pressure, 288
Weight, effect of thyroid hormone on, 901-902
  gain, during pregnancy, 990
  loss, in diabetes, 934

'eightlessness, in space, 536
'ernicke's aphasia, 662
'ernicke's area, 654-656, 662
'hite blood cells. See *Leukocytes.*
:hite pulp, of spleen, 343
'hite ramus, 687
'indkessel, 162
'ithdrawal reflexes, 613-614
'ord blindness, 654, 662, 728
'ord deafness, 662
'ork capacity, at high altitudes, 532
'ork output. See *Cardiac output*

X chromosome, 955, 983
Xenografts, 74

Y chromosome, 955, 983
"Y" proteins, 838
Yield pressure, 214
Young-Helmholtz theory, 718

Z disc, 121
"Z" proteins, 838
Zero reference point, 188
Zero reference potential, 193
Zinc, metabolic functions of, 872
Zona fasciculata, 909-910
Zona glomerulosa, 909-910, 913
Zona pellucida, 983-984
Zona reticularis, 909-910
Zonula occludens, 401
Zygote, 984
implantation of, 984-985