CHAPTER 51

Quality Assurance and Control

John H Parker, PhD
John E Enders, PhD, MBA

The pharmaceutical industry, as a vital segment of the healthcare system, conducts research and manufactures and markets pharmaceutical and biological products and medical devices used for the acute/chronic treatment and diagnosis of disease. Recent advances in drug discovery, primarily in the field of biotechnology and in the required controls over manufacturing processes, are presenting new challenges to the control of quality and to the systems that operate internally in the industry. The external regulations established by the federal Food and Drug Administration (FDA) and other regulatory bodies also add to these challenges. The evolving role of the industrial quality professional requires more extensive education including food and drug law, business, as well as the traditional science/technology coursework.

The pursuit of quality is being approached through the application of quality systems such as Total Quality Management (TQM) and continuous improvement, whereby management and labor join forces to build quality into products while helping to ensure the company's financial success. This changed emphasis is directed toward defect prevention (proactive) rather than defect detection (after the fact).

Quality assurance (QA) and quality control (QC) groups develop and follow standard internal operating procedures directed toward assuring the quality, safety, purity, and effectiveness of drug products. The FDA has issued a primary regulation to the industry entitled *Current Good Manufacturing Practice for Finished Pharmaceuticals* (commonly referred to as the cGMPs or GMPs). Numerous guidelines have been issued relative to specific dosage forms and operations such as aseptic manufacturing, validation and stability testing, etc., which impose significant compliance requirements. These guidelines also serve as the basis for compliance investigations conducted by the FDA and are used in regulatory agency inspections of facilities and operations. Recently, emphasis is being placed on the inspection of quality systems as part of the regulatory pre-approval program when reviewing submissions relative to New Drug Applications (NDAs) and Biological License Applications (BLAs).

QA AND QC: ORGANIZATION/ RESPONSIBILITIES

Industry, to ensure compliance with these government regulations and with their own internal policies and procedures, has developed very sophisticated quality organizations with well-defined responsibilities. It has been accepted that QA and QC have different functions within an organization; although both are considered part of the Quality Unit as identified in 21CFR. QC most commonly functions to test and measure material and product. QA establishes systems for ensuring the quality of the product. Firms must decide upon the exact roles they wish QC and QA to perform in operations and put these definitions in writing.

QA Functions and Responsibilities

The QA department within any organization, because of its responsibilities, normally will report to a relatively high-level administrator within a company, depending on its size. In smaller companies they may report to the chief executive officer or the president. In larger corporations they will sometimes report to the president or executive vice-president or chief of operations. In any case, however, responsibility for quality, as currently dictated by FDA, ultimately resides with top management.

In all cases QA will be independent of the economic issues associated with manufacturing and distribution of the product. The QA department is responsible for ensuring that the quality policies adopted by a company are followed. In some organizations, QA serves as the primary contact with regulatory agencies and is the final authority for product acceptance (release) or rejection. It is customary for QA to play a major role in the identification and preparation of the necessary policies and standard operating procedures (SOPs) relative to the control of quality. Where it has responsibility for final product release, it must determine that the product meets all the applicable specifications and that it was manufactured according to internal standards and cGMPs. QA departments now tend to work as a team member with the other functional groups within the firm rather than simply to serve a police function, a largely outdated role of QA.

A second major responsibility of the QA department is the quality monitoring or audit function. Through this activity it is able to determine if operations have adequate systems, facilities, and written procedures to control the quality of products produced. Thus, the QA function not only determines that the procedures are current and correct, but that properly trained operators are following them. Combining this review of SOPs with an audit of facilities and operations will give company management an inside report on its level of compliance and will allow the necessary changes and/or corrections to be made prior to either causing a product failure or being reported as a deficiency during an inspection by an FDA investigator. This is consistent with the top-level management review component of the quality systems approach currently emphasized by FDA during inspections.

Senior management of a company looks to the QA function to assess operations continually and to advise and guide them toward full compliance with all applicable internal and external regulations. Organizationally, the Quality Department(s) should report, as directed by the GMPs, to someone other than the person responsible for production.

QC Functions and Responsibilities

Quality Control is responsible for the day-to-day control of quality within a company. This department is staffed with scientists and technicians responsible for the sampling and analytical testing of incoming raw materials and inspection of packaging components, including labeling. They conduct in-process testing when required, perform environmental monitoring, and inspect operations for compliance. Finally, they conduct the required tests on finished dosage form. QC is also responsible for monitoring product quality through distribution, including testing of product complaint samples, evaluating product stability, etc.

Many companies have the heads of QC and production report to a common higher level of management, but with QC being independent of production. This higher-level management may be the same or different individuals, but it allows independent operation of both functions without direct conflict arising when reaching a final decision on the acceptability of final products.

The analytical control laboratory must be staffed with persons who are trained academically and are, through experience, capable of performing the often complex analyses used to evaluate the acceptability of the materials tested. The equipment and instrumentation in the laboratory must be suitable for performing the testing in an accurate and efficient manner. Detailed specifications must be available, as well as validated test methods against which products and raw materials will be evaluated. The specifications detail the limits for acceptance of the article, based on identified critical parameters.

The testing and acceptance of only high-quality raw materials is essential for the production of uniformly acceptable products. Quality Control plays a major role in the selection and qualification of vendors from whom these materials are purchased. Testing of representative samples is required, and in many cases, an audit of the vendor's operation is necessary to determine their suitability and degree of compliance with GMPs and other relevant standards prior to their being approved. The vendor audit frequently is organized by QA, with technical support from research, QC, and manufacturing.

At various critical in-process steps in production, it may be necessary to sample and test product against criteria previously established. Often, in-process alert or action levels will be identified for the critical in-process parameters as a means of process control. These alert or action levels are normally set such that they are more restrictive than the final acceptance limits, but serve as an in-process control by providing early warnings of conditions that could lead to an out-of-control situation and thus will allow timely corrective action before such conditions occur. Trending of analytical data is also useful in providing early warning signals that the process is moving out of control. It should be noted, however, that materials, which have reached the alert or action level criteria, are still acceptable for use in manufacturing, since they have not exceeded an out-of-limit rejection level.

Quality Control is responsible, as part of its testing and inspection functions, for monitoring the environmental conditions under which products are manufactured and/or held. Different levels of control are established depending on the intended use of the dosage form. Parenteral and ophthalmic products must be produced in a controlled environment that is designed to ensure their sterility. Monitoring of air and water systems is critical in confirming that they are being controlled and that the levels of particulates, microbial matter, and other contaminants are within pre-established limits. The USP contains monographs and specifications on Water Used for Pharmaceutical Purposes. Formerly, the Federal Government Standard 209E, *Airborne Particulate Cleanliness Classes in Cleanrooms and Clean Zones,* established acceptable limits for particulates in a controlled environment, but is no longer considered applicable to the pharmaceutical industry. Federal standards are currently not enforced for environmental quality, but guidance is available in the FDA Concept Paper, Sterile Drug Products Produced by Aseptic Processing–Draft, published on September 27, 2002. In addition, reference is made to the Baseline Pharmaceutical Engineering Guide, Vol. 3, Sterile Manufacturing Facilities, published by the International Society of Pharmaceutical Engineering (ISPE) in partnership with the FDA, in June 2000. Generally, conditions listed as Class 100 (or equivalent) are maintained in areas where parenteral products are filled into clean, sterile containers. Class 100 is defined as an area that can be controlled to contain fewer than 100 particles, 0.5 μm and larger, per cubic foot of air. In addition, manufacturers must establish limits for the presence of viable microorganisms in the environment and appropriately monitor the air quality in the filling area.

Another major element of control is the environmental monitoring of the areas in which nonsterile products are manufactured, such as liquids, tablets, and capsules. The objective here is first to determine an acceptable level of particulates and microbial contaminates and then to control them to this level. If particulate levels are found to be excessive, steps must be taken to bring them within acceptable limits so as not to compromise the quality of the product. This monitoring and control of the environment will further ensure the quality and stability of the product by preventing the products from being exposed to a hostile environment.

Control of packaging components, especially those that come into direct contact with a product, is required. These materials must be inspected and tested against rigid specifications to ensure that they meet predetermined functional standards. This includes evaluation of compatibility of the product with the packaging materials. Labeling is understandably a critical component, not just in its original design and acceptance, but also with regard to secure storage and issuance to ensure accountability. Furthermore, final product labeling must be 100% inspected to ensure that it is correct.

TOTAL QUALITY MANAGEMENT/TOTAL PRODUCT QUALITY

The production of quality pharmaceutical products requires embracing the principles of Total Quality Management (TQM). Although the term TQM has fallen out of favor in recent years and has been replaced by other, though similar designations, such as Total Product Quality (TPQ), the principles of TQM will serve to improve productivity and customer satisfaction. The quality function is part of a team composed of research, production, marketing/sales, and customer service. In the competitive markets of today, it is critical to improve quality and service continually while minimizing costs and maximizing resource utilization to help contain overall health care costs. The concept of TQM requires the total commitment of senior-level management and supervision of all departments, operators, suppliers, and customers. Its basic principle is one of continually striving for process improvement that begins with product development and only concludes when feedback and follow-up have been completed on customer complaints and suggestions. In many firms the QA Department has the responsibility to organize and implement programs with these objectives in mind.

Quality must be designed into products, beginning with research and development phases. Product quality criteria are established, and detailed specifications are written. Meticu-

lous, written procedures must be prepared for production and control, and processes must be rigorously validated. Raw materials must be characterized and then purchased from reputable, approved suppliers to ensure that, when the materials are incorporated into the finished dosage form, they will provide products of uniformly high quality. Facilities must be designed, constructed, and controlled to provide the proper stable environment for protecting the integrity of products. Equipment must be selected that is efficient and can be cleaned readily and sanitized, to aid in preventing cross- contamination of one product with another. Personnel must be trained properly so that their personal habits, clothing, and job performance do not compromise product quality. The directions that they use must be in writing, approved by responsible individuals, and strictly followed. Training programs must be thoroughly documented and include an evaluation of mastery of the procedures employed.

Distribution departments are responsible for controlling the shipping and handling of products, using inventory-control systems based on the *first in–first out* principle. They select modes of distribution that will protect products from adverse handling or environmental conditions while in transit to distribution points and to customers. Furthermore, they must maintain accurate records of distribution to ensure that any product recall, if required, will be effectively and thoroughly conducted.

The marketing department must be sensitive to the customer's needs and be responsive to complaints. The quality department should be kept informed of real or potential problems as reported from the marketplace so that they may conduct investigations of product complaints, as appropriate, to determine the cause of the condition described in the complaint.

Involved with each of the operations described above, QA is ever present and gives approval only after assessing and being assured that the entire production process has been completed satisfactorily and that all the aspects of the GMPs have been satisfied.

In the pharmaceutical industry, TQM or the equivalent system, therefore, can be looked upon as a combined team effort to develop, produce, market, distribute, and control products that are safe and will be effective for the time they remain in the marketplace. Such a program ultimately will assure the professional dispenser and the final consumer that each lot of every product conforms to certain specifications and that each unit has met all the requirements, both internal and external, of the industry and will fulfill the declarations made in its labeling.

DOCUMENTATION

The saying, "If it wasn't documented, it wasn't done," describes the linkage between written records of action taken and the quality operation. These written documents include those found in the product-development phase as well as those associated with the actual manufacture and testing of individual batches. The former will consist of research and development reports, technology transfer reports, and the validation records required when the FDA conducts its pre-approval investigations. Elements of these documents will include raw material and final product specifications along with appropriate validated test methods, technology-transfer documents, and production scale-up support data. Specific critical pieces of equipment must be identified along with the process and product qualification/validation records. The *Master Production Batch Record* (MPBR) is often the document that facilitates the orderly transition from product and process development to commercial-scale production, since it is the document that captures the process as described by the product development documentation.

The *Production Batch Record* (PBR), an exact copy of the approved MPBR, is used along with written SOPs to produce individual batches of product that are assigned specific code or lot numbers. The PBR provides a historical record of every step, beginning with the receipt of raw materials and package components and continuing through each phase of production. Recording charts or computer printouts of significant operations such as autoclaving, drying, air-particulate monitoring, lyophilizing, etc, become part of the batch history. After a batch has been completed, including final analytical and physical testing, there is one final additional step that must be completed prior to approving the batch for distribution. All documentation relating to the production of the specific batch is given a final review. This is often a two-step process in which each of the required documents is checked for accuracy and completeness by manufacturing. Any discrepancy must then be investigated and a written explanation made. This is followed by a final review by QA to ensure that all documents are complete and that all issues have been appropriately resolved. Only after this final review by QA has been completed satisfactorily may the batch be approved for release. Once the batch has been approved, accurate distribution records are required, to trace the batch in the marketplace, which would facilitate, if the need arose, recalling the batch.

QUALITY IN PHARMACEUTICAL BIOTECHNOLOGY

Because of the physical/chemical nature of the proteinaceous products derived from pharmaceutical biotechnology, unique quality considerations prevail that are associated with early research and development synthesis, clinical product scale-up, and commercial manufacture. The reason that this is particularly challenging becomes more evident when some gross differences between biological therapeutic agents and chemical drug products are examined. In contrast to small-molecule pharmaceuticals, biotechnology-derived drugs are obtained from living organisms and often consist of complex mixtures of protein and other substances, often heat labile, and, finally, highly susceptible to microbiological contamination. In the context of this discussion, pharmaceutical biotechnology products are defined to include proteins and peptides produced by recombinant DNA (rDNA) techniques and monoclonal antibody/hybridoma (Mab) technology. The former refers to the fact that these products are often produced by microorganisms or mammalian cells containing hybrid DNA, most often produced by joining pieces of DNA from different sources to gain the appropriate expressed product; the latter involves production of a single clone from hybrid cells, using hybridoma technology that fuses different cells to make the desired antibody. See also Chapter 49, *Biotechnology and Drugs*. In general, the object of the pharmaceutical biotechnology manufacturing process is to produce a product essentially free of contamination. The attributes that the product should possess are sterility and the absence of pyrogens, unwanted organisms, by-products of the manufacturing process, and degradation by-products. The same requirements for adherence to cGMPs exist throughout the production process for recombinant proteins and monoclonal antibodies as apply to other pharmaceutical products.

Consideration must also be given to the design of the delivery system for the biotechnology-derived drug, especially because of the lability of many of these products. In accomplishing this task, often the more conventional manufacturing processes may be employed such as sterile filtration, aseptic handling, and, in some instances, lyophilization. However, for many of these products, new delivery systems are being evaluated as described below.

Characterization of the products produced by pharmaceutical biotechnology is a rapidly changing technical challenge, but tremendous advances have been made in the past few years. The FDA considers many of these products to be well characterized and has changed the approach to license approval with this improved understanding of product specifications in mind. Bogdansky[1] outlines QC or testing considerations that address

the structure, potency, and purity of proteins and the analysis of contaminants resulting from the manufacturing process or degradation. Full characterization includes physical and chemical stability. Satisfactory stability of the product is a requirement for controlled manufacture and an acceptable shelf life following distribution.

CONTEMPORARY ISSUES

As a demonstration of the pace of change in the pharmaceutical industry, technologies that were on the cutting edge as recently as 5 years ago have become rather commonplace, as organizations routinely employ them to increase productivity and reduce costs while maintaining product quality.

Statistical process control permits improved real-time control, thereby reducing end-product failures. Qualification and potential certification of suppliers of goods and services adds to the thrust of building in quality and allows the reduction in inventory costs by following just-in-time purchasing and receipt principles. Finally, in practically all facets of research, development, and operations, automation and computerization, including robotics, are work methods that have an impact on our daily lives by increasing productivity and raising the standards of quality by enhancing reproducibility. This application of computerized manufacturing and control has led to increasingly rigorous requirements for the appropriate design and validation of computer-controlled processes. This is evident in the publishing of 21CFR part 11 and GAMP4, which describe the requirements for electronic records and signatures and validation respectively. It should be noted, however, that in January 2003, draft guidance from FDA on these matters was withdrawn, accompanied by a statement from FDA that the approach to these issues is changing. New FDA guidance is expected to be forthcoming in the not-too-distant future.

Today, the pace of change in the pharmaceutical industry continues unabated. There is increased emphasis on analytical chemistry as it relates to the entire drug discovery, development, and manufacturing sequence. This, coupled with new clues as to potential drug targets uncovered through the Human Genome Project and new approaches to computer modeling of potential drug compounds, has led to many new candidates for development. This is evident in the new drug submissions made by industry to the FDA and in the depth and complexity of the review process associated with these new compounds. More technically sophisticated instrumental methods of analysis, assisted by computer interfacing, provide greater sensitivity along with the ability to analyze the results more efficiently and effectively. From these advances flows the requirement for more stability-indicating assay methods as well as increased emphasis on the impurities in drug substances and products, such as organic volatile impurities and even ordinary impurities. Taken as a whole, accurate mass balance of the parent compound, degradents, and impurities is an expectation. Compendia such as the USP are including these concepts in general chapters and individual drug monographs. The evolution of high-pressure liquid chromatography (HPLC) methods, and other even more sensitive technologies, and the wide acceptance of these techniques has increased the trend toward painstakingly thorough characterization of products, so that it is becoming the exception rather than the rule. Additionally, HPLC facilitates the recent focus on optical purity through chiral separation to support improvements in asymmetric synthesis, with the intent of producing the single, therapeutically active compound.

The move toward the elimination of tests that require animals is exemplified by the replacement of the rabbit pyrogen test with the bacterial endotoxin (LAL) method. It may be expected that similar chemical and biochemical approaches will be developed to eliminate the use of animal testing in pre-clinical drug evaluation and toxicological testing as well.

The concept of parametric release of end-product is being applied on the basis of complete knowledge and control of all phases of the process including such things as information on suppliers, process and product validation, operator training, and thorough process knowledge coupled with statistical process controls. The sum of these prospectively managed and controlled quality activities results in greater real-time control and, hence, a diminished need for end-product confirmatory testing.

In the operations area, newer types of dosage forms, such as liposomes and transdermal devices, are demanding innovative manufacturing and control procedures and practices. Novel routes of administration such as pulmonary inhalation for the administration of insulin are also being explored. Automation and computerization continue to increase manufacturing yields and concomitant tighter tolerances. As a result, there is a renewed interest in functional testing of raw materials, to keep up with these manufacturing advances.

All of these advances will, no doubt, be refined with the passage of time and continuing diligent efforts on the part of industrial professionals and regulators. This will lead to ever-new issues affecting the industrial quality professional and, hence, the challenge and reward of such an exciting endeavor.

FDA MODERNIZATION

The FDA Modernization Act (FDAMA) of 1997 was introduced with the intent of improving the review process and thereby making new drug products available to the public faster than in the past. This effort continues both in form and detail and, in general, has shortened the review and approval timeline. A recent decision by FDA to merge many of the product categories regulated by the Center for Biologics (CBER) into the Center for Drugs (CDER) is intended to further expedite the review process. See Chapter 48 for more information on the new drug approval process.

GMP REGULATIONS

In March 1979, the FDA issued revised GMP regulations. These regulations, still in effect today, present the minimum requirements to be met by industry when manufacturing, packaging, and holding human and veterinary drugs.

The FD&C Act states that a drug is deemed to be adulterated unless the methods used in its manufacture, processing, packing, and holding, as well as the facilities in which it was produced and the controls used during its production, conform to the GMPs so that the drug will meet the safety requirements of the Act and that it has the correct identity and strength to meet the quality and purity characteristics that it is represented to possess. Through the intervening years additional regulations and guidelines have been issued to supplement the original drug GMPs such as those for the *Good Manufacturing Practice for the Manufacture, Packing, Storage and Installation of Medical Devices* and *Good Laboratory Practice (GLPs) for Controlling and Conducting Human Clinical Studies*. Additionally, guidelines have been issued relating to the *Manufacture and Control of Large Volume Parenteral Solutions, Control of Sterile Products Produced by Aseptic Processing, Inspections of Bulk Pharmaceutical Chemicals,* and an *Inspection Guide for Quality Control Laboratories,* as well as on many other topics. A number of other guidelines or concept papers have been prepared by various organizations within the industry itself, such as the Pharmaceutical Manufacturers Association (PhRMA), the Parenteral Drug Association (PDA), the International Society of Pharmaceutical Engineering (ISPE), and others. A partial listing is provided in the *Bibliography*.

The current GMP regulations and these additional guides and guidelines should be read and understood thoroughly by

those involved in or interested in pursuing QC and QA responsibilities. The scope of the present regulation is given in the following outline, along with a brief interpretation of each subpart.

REFERENCES

1. Bogdansky FM. *Pharm Technol* 1987; (Sep): 72.

BIBLIOGRAPHY

Human and Veterinary Drugs—Current Good Manufacturing Practice for Finished Pharmaceuticals. 21 CFR 211: 2002.

Airborne Particulate Cleanliness Classes in Cleanrooms and Clean Zones (Fed Std 209E). Washington, DC: GSA, 1992.

Quality System Regulation in the Manufacturing of Medical Devices. 21 CFR 820: 2002.

Good Laboratory Practice for Non-clinical Laboratory Studies. 21 CFR 58: 2002. Rockville, MD: USP/NF, USPC.

Validation of Steam Sterilization Cycles (Tech Monogr #1). Philadelphia: PDA, 1978.

Validation of Dry Heat Processes Used for Sterilization and Depyrogenation (Tech Monogr #3). Philadelphia: PDA, 1981.

Sterile Pharmaceutical Packaging: Compatibility and Stability (Tech Rep #5). Philadelphia: PDA, 1984.

Fundamentals of a Microbiological Environment Monitoring Program (Tech Rep #13). Philadelphia: PDA, 1990.

Current Practices in the Validation of Aseptic Processing—1992 (Tech Rep #17). Philadelphia: PDA, 1993.

Process Simulation Testing for Aseptically Filled Products-1996 (Tech Rep # 22). Philadelphia: PDA, 1996.

Technical Report: Process Simulation Testing for Aseptically Filled Products (Tech Rep #22). Philadelphia: PDA, 1996.

Points to Consider for Aseptic Processing–2003 (Supplement Volume 57, #2). Philadelphia: PDA, 2003

Guideline on Sterile Products Produced by Aseptic Processing. Rockville, MD: FDA, Jun 1987.

Sterile Drug Products Produced by Aseptic Processing–Draft. Rockville, MD: FDA, Sep 2002.

Concepts and Principles for the Validation of Computer Systems Used in the Manufacture and Control of Drug Products. *Proc PMA Sem* Apr 1986.

Risk-Based Approach to 21 CFR Part 11–ISPE White Paper. Tampa, FL: ISPE, 2003.

Points to Consider in the Manufacture and Testing of Monoclonal Antibody Products for Human Use. Rockville, MD: FDA, Feb 1997.

Points to Consider on Plasmid DNA Vaccines for Preventive Infectious Disease Indications. Rockville, MD: FDA, Dec 1996.

Huxsoll JF. *Quality Assurance for Biopharmaceuticals*. New York: Wiley, 1994.

Guide to Inspection of Bulk Pharmaceutical Chemicals. Rockville, MD: FDA, Sep 1992.

FDA Website [http://www.fda.gov] has many current guidance documents available for review and/or downloading.

GMP Training Organizations Website [http://gmptraining.com/news.html] has links to several organizations including DIA, ISPE, and PDA for current information on quality issues.

Other Websites providing useful references: [http://www.ispe.org/; http://www.diahome.org; http://www.raps.org/; http://www.pda. org/;]

Current Good Manufacturing Practices

APPENDIX A

CFR Title 21 Food and Drugs

PART 211 CURRENT GOOD MANUFACTURING PRACTICE FOR FINISHED PHARMACEUTICALS

SUBPART A GENERAL PROVISIONS

211.3 (Definitions) The scope of the regulations are explained for human prescription and OTC drug products including drugs used to produce medicated animal feed. Reference is made to Part 210.3 of the chapter that gives definitions for all significant terms used in the regulations.

SUBPART B ORGANIZATION AND PERSONNEL

211.22 (Responsibilities of QC unit) Highlighted here is the assignment to the QC unit of total responsibility for ensuring that adequate systems and procedures exist and are followed to ensure product quality.

211.25 (Personnel qualifications) Personnel, either supervisory or operational, must be qualified by training and experience to perform their assigned tasks.

211.28 (Personnel responsibilities) The obligations of personnel engaged in the manufacture of drug products concerning their personal hygiene, clothing, and medical status are defined.

211.34 (Consultants) The qualifications of consultants must be sufficient for the project to which they are assigned.

SUBPART C BUILDINGS AND FACILITIES

Buildings and facilities can be considered acceptable only if they are suitable for their intended purpose and can be maintained. Construction concepts, such as air handling systems, lighting, eating facilities, and plumbing systems including water, sewage and toilet facilities, are outlined.

211.42 (Design and construction features)

211.44 (Lighting)

211.46 (Ventilation, air filtration, air heating and cooling)

211.48 (Plumbing)

211.50 (Sewage and refuse)

211.52 (Washing and toilet facilities)

211.56 (Sanitation)

211.58 (Maintenance)

SUBPART D EQUIPMENT

Equipment must be designed, constructed, of adequate size, suitably located, and able to be maintained and cleaned to be considered suitable for its intended use. Reference is made to the use of automatic equipment, data processors, and computers, highlighting the need for input/output verification and for proper calibration of recorders, counters, and other electrical or mechanical devices.

211.63 (Equipment design, size, and location)

211.65 (Equipment construction)

211.67 (Equipment cleaning and maintenance)

211.68 (Automatic, mechanical, and electronic equipment)

211.72 (Filters) Special note is made that the only filters to be used are those that do not release fibers into products.

SUBPART E CONTROL OF COMPONENTS AND DRUG PRODUCT CONTAINERS AND CLOSURES

211.80 (General requirements) Written procedures must be available that describe the receipt, identification, storage, handling, sampling, testing, and approval or rejection of components (raw materials) and drug products.

211.82 (Receipt and storage of untested components, drug product containers, and closures)

211.84 (Testing and approval or rejection of components, drug product containers, and closures)

211.86 (Use of approved components, drug product containers, and closures) These shall be rotated so that the oldest approved stock is used first.

211.87 (Retesting of approved components, drug product containers, and closures) Materials that are subject to deterioration during storage should be retested at an appropriate time based on stability profiles.

211.89 (Rejected components, drug product containers, and closures) These shall be identified and controlled to prevent their use in manufacturing.

211.94 (Drug product containers and closures) Containers and closures (product contact materials) must protect the product and must be nonreactive with or additive to the product, suitable for their intended use, and controlled using written procedures.

SUBPART F PRODUCTION AND PROCESS CONTROLS

211.100 (Written procedures; deviations) Written standard operating procedures (SOPs) for each production process and control procedure are necessary. Any deviation from an SOP must be investigated, recorded, and approved prior to final product acceptance.

211.101 (Charge-in of components) The procedures used to formulate a batch shall be written and followed.

211.103 (Calculation of yield) Actual yields and theoretical yields shall be determined. All products are to be formulated to provide not less than 100% of the required amount of active ingredient. Records are to be maintained of each component and the quantity, which is incorporated into a batch.

211.105 (Equipment identification) Equipment shall be properly identified.

211.110 (Sampling and testing of in-process materials and drug products) Significant in-process steps are to be identified and appropriate sampling, testing, and approvals obtained before proceeding further in the production cycle. Rejected material must be controlled.

211.111 (Time limitations on production) If required, time limitations will be placed on in-process steps.

211.113 (Control of microbiological contamination) Appropriate procedures are to be prepared for the control and prevention of microbiological contamination. The sterilization process must be validated.

211.115 (Reprocessing) Reprocessing of product is allowed providing there are written procedures covering the methods and QC unit review to be used.

SUBPART G PACKAGING AND LABELING CONTROL

211.122 (Materials examination and usage criteria) Labeling and packaging materials are to be received, identified, stored, sampled, and tested following detailed written procedures.

211.125 (Labeling issuance) Strict control shall be exercised over labeling for use in drug product labeling operations

211.130 (Packaging and labeling operations) There shall be written procedures designed to ensure that correct labels, labeling, and packaging materials are used for drug products. Special controls must be exercised over labeling to ensure that only the correct labels are issued to packaging for a specific product and that the quantities used are reconciled with the quantity issued.

211.132 (Tamper-resistant packaging requirements for over-the-counter (OTC) human drug products) Provides details of tamper-resistant packaging.

211.134 (Drug product inspection) Packaged and labeled products shall be inspected for correct labels.

211.137 (Expiration dating) Following appropriate stability studies at prescribed temperature conditions, products on the market shall bear an expiration date to ensure that they are used within their expected shelf life.

SUBPART H HOLDING AND DISTRIBUTION

211.142 (Warehousing procedures) Describes the requirements for warehousing holding product under appropriate conditions of light, temperature, and humidity.

211.150 (Distribution procedures) Written procedures describing product distribution shall be prepared

SUBPART I LABORATORY CONTROLS

211.160 (General requirements) Describes the general requirements for laboratory control mechanisms.

211.165 (Testing and release for distribution) Concerns written procedures in the form of specifications, standards, sampling plans, and test procedures that are used in a laboratory for controlling components and finished drug products. Acceptance criteria for sampling and approval shall be adequate to support release of product for distribution.

211.166 (Stability testing) There shall be a written testing program designed to assess the stability characteristics of drug products. The results of this testing shall be used in assigning appropriate storage conditions and expiration dates.

211.167 (Special testing requirements) Special testing requirements are given for sterile and/or pyrogen-free ophthalmic ointment and controlled-release dosage form products.

211.170 (Reserve samples) Reserve sample quantity and retention times are described.

211.173 (Laboratory animals) Animals used in any testing shall be maintained and controlled in a manner suitable for use.

211.176 (Penicillin contamination) Drug products cannot be marketed if, when tested by a prescribed procedure, found to contain any detectable levels of penicillin.

SUBPART J RECORDS AND REPORTS

211.180 (General requirements) Describes record retention time and availability for inspection.

211.182 (Equipment cleaning and use log) A written record of major equipment cleaning, maintenance, and use shall be included in major equipment logs.

211.184 (Component, drug product container, closure, and labeling records) Deals with the issues of the receipt, testing, and storage of components, drug product containers, and closures. Details the various records and documents that should be generated during the manufacture of drug products and that are to be available for review.

211.186 (Master production and control records) A master production record must be prepared for each drug product, describing all aspects of its manufacture, packaging, and control. Individual batch records are derived from this approved master.

211.188 (Batch production and control records) Calls for batch production and control records with information about the production and control of each batch

211.192 (Production record review) All drug product batch records shall be reviewed and approved by the QC unit (QA/QC) before the batch is released.

211.194 (Laboratory records) Complete records of any laboratory testing shall be maintained to include raw data, test procedures and results, equipment calibration, and stability testing.

211.196 (Distribution records) Distribution records include warehouse shipping logs, invoices, bills of lading, and all documents associated with distribution. These records should provide all the information necessary to trace lot distribution to facilitate product retrieval if necessary.

211.198 (Complaint files) Records of complaints received from consumers and professionals are to be maintained along with the report of their investigation and response.

SUBPART K RETURNED AND SALVAGED DRUG PRODUCTS

211.204 (Returned drug products) Records are to be maintained of drug products returned from distribution channels and the reason for their return. These data can be used as part of the total lot accountability, should the need arise, to trace its distribution and/or for its recall.

211.208 (Drug product salvaging) Drug products that have been stored improperly are not to be salvaged.

CHAPTER 52

Stability of Pharmaceutical Products

Patrick B O'Donnell, PhD
Allan D Bokser, PhD

Stability of a pharmaceutical product may be defined as the capability of a particular formulation, in a specific container/closure system, to remain within its physical, chemical, microbiological, therapeutic, and toxicological specifications. Assurances that the packaged product will be stable for its anticipated shelf life must come from an accumulation of valid data on the drug in its commercial package. These stability data involve selected parameters that, taken together, form the stability profile. Pharmaceutical products are expected to meet their specifications for identity, purity, quality, and strength throughout their defined storage period at specific storage conditions.

The stability of a pharmaceutical product is investigated throughout the various stages of the development process. The stability of a drug substance is first assessed in the preformulation stage. At this stage, pharmaceutical scientists determine the drug substance and its related salts stability/compatibility with various solvents, buffered solutions, and excipients considered for formulation development. Optimization of a stable formulation of a pharmaceutical product is built upon the information obtained from the preformulation stage and continues during the formulation development stages.

Typically, the first formulation development stage is the preparation of a "first in human" formulation which is often a non-elegant formulation optimized for short-term dose-ranging clinical studies. The second major formulation development stage occurs to support Phase II and early Phase III clinical studies. The pharmaceutical product developed at this stage is usually the prototype for the commercial product. Therefore, the pharmaceutical product will be formulated based in part on the stability information obtain from the previous formulations and must meet stability requirements for longer-term clinical studies. The final formulation development stage is for the commercial pharmaceutical product. In addition to building on the clinical requirements of the drug, the commercial pharmaceutical product must also incorporate the commercial or the final market image of the product, which includes the container closure system. The stability of this product must be demonstrated to the appropriate regulatory agencies in order to assign an expiration date for the product.

Once a pharmaceutical product has gained regulatory approval and is marketed, the pharmacist must understand the proper storage and handling of the drug. In some cases, a pharmacist may need to prepare stable compounded preparations from this product. It is the responsibility of the pharmacist, via the information of the manufacturer, to instruct the patient in the proper storage and handling of the drug product. The impact of a drug product with a poor stability profile could delay approval, affect the safety and efficacy of the drug, and/or cause product recall.

Much has been written about the development of a stable pharmaceutical product. Comprehensive treatments of all aspects of pharmaceutical product stability has been published by Lintner,[1] Connors et al,[2] and more recently Carstensen[3]. This chapter will outline the appropriate steps from preformulation to drug approval to assure that the pharmaceutical product developed is stable. Requirements for compounded products will also be discussed.

The USP defines the stability of a pharmaceutical product as "extent to which a product retains, within specified limits, and throughout its period of storage and use (ie, its shelf-life), the same properties and characteristics that it possessed at the time of its manufacture." There are five types of stability that must be considered for each drug.

Type of Stability	Conditions Maintained Throughout the Shelf-Life of the Drug Product
Chemical	Each active ingredient retains its chemical integrity and labeled potency, within the specified limits.
Physical	The original physical properties, including appearance, palatability, uniformity, dissolution, and suspendability are retained.
Microbiological	Sterility or resistance to microbial growth is retained according to the specified requirements. Antimicrobial agents that are present retain effectiveness within the specified limits.
Therapeutic	The therapeutic effect remains unchanged.
Toxicological	No significant increase in toxicity occurs.

Stability of a drug also can be defined as the time from the date of manufacture and packaging of the formulation until its chemical or biological activity is not less than a predetermined level of labeled potency and its physical characteristics have not changed appreciably or deleteriously. Although there are exceptions, 90% of labeled potency generally is recognized as the minimum acceptable potency level. Expiration dating is defined, therefore, as the time in which a drug product in a specific packaging configuration will remain stable when stored under recommended conditions.

An expiration date, which is expressed traditionally in terms of month and year, denotes the last day of the month. The expiration date should appear on the immediate container and the outer retail package. However, when single-dose containers are packaged in individual cartons, the expiration date may be

placed on the individual carton instead of the immediate product container. If a dry product is to be reconstituted at the time of dispensing, expiration dates are assigned to both the dry mixture and the reconstituted product. Tamper-resistant packaging is to be used where applicable.

One type of time-related stability failure is a decrease in therapeutic activity of the preparation to below labeled content. A second type of stability failure is the appearance of a toxic substance, formed as a degradation product upon storage of the formulation. The numbers of published cases reflecting this second type are few. However, it is possible, though remote, for both types of stability failures to occur simultaneously within the same pharmaceutical product. Thus, the use of stability studies with the resulting application of expiration dating to pharmaceuticals is an attempt to predict the approximate time at which the probability of occurrence of a stability failure may reach an intolerable level. This estimate is subject to the usual Type 1 or alpha error (setting the expiration too early so that the product will be destroyed or recalled from the market appreciably earlier than actually is necessary) and the Type 2 or beta error (setting the date too late so that the failure occurs in an unacceptably large proportion of cases). Thus, it is obligatory that the manufacturer clearly and succinctly define the method for determining the degree of change in a formulation and the statistical approach to be used in making the shelf-life prediction. An intrinsic part of the statistical methodology must be the statements of value for the two types of error. For the safety of the patient a Type 1 error can be accepted, but not a Type 2 error.

REGULATORY REQUIREMENTS

Stability study requirements and expiration dating are covered in the Current Good Manufacturing Practices (cGMPs),[4] the USP,[5] and the FDA guidelines.[6]

GOOD MANUFACTURING PRACTICES—The GMPs[4] state that there shall be a written testing program designed to assess the stability characteristics of drug products. The results of such stability testing shall be used to determine appropriate storage conditions and expiration dating. The latter is to ensure that the pharmaceutical product meets applicable standards of identity, strength, quality, and purity at time of use. These regulations, which apply to both human and veterinary drugs, are updated periodically in light of current knowledge and technology.

COMPENDIA—The compendia also contain extensive stability and expiration dating information. Included are a discussion of stability considerations in dispensing practices and the responsibilities of both the pharmaceutical manufacturer and the dispensing pharmacist. It now is required that product labeling of official articles provide recommended storage conditions and an expiration date assigned to the specific formulation and package. Official storage conditions as defined by the USP 26[5] are as follows: *Cold* is any temperature not exceeding 8°C, and *refrigerator* is a cold place where the temperature is maintained thermostatically between 2 and 8°C. A *freezer* is a cold place maintained between −25 and −10°C. *Cool* is defined as any temperature between 8 and 15°C, and *room temperature* is that temperature prevailing in a working area. *Controlled room temperature* is that temperature maintained thermostatically between 20 and 25°C. *Warm* is any temperature between 30 and 40°C, while *excessive heat* is any heat above 40°C. Should freezing subject a product to a loss of potency or to destructive alteration of the dosage form, the container label should bear appropriate instructions to protect the product from freezing. When no specific storage instructions are given in a USP monograph, it is understood that the product's storage conditions shall include protection from moisture, freezing, and excessive heat.

As is noted above in USP 26, the definition of controlled room temperature was a "temperature maintained thermostatically between 20 and 25°C (68 and 77°F)." This definition was established to harmonize with international drug standards efforts. The usual or customary temperature range is identified as 20 to 25°C, with the possibility of encountering excursions in the 15 to 30°C range and with the introduction the mean kinetic temperature (MKT).

The mean kinetic temperature is calculated using the following equation:

$$T_k = \left[\frac{\Delta H/R}{-In\left(\frac{e^{-\Delta H/RT_1} + e^{-\Delta H/RT_2} + \ldots + e^{-\Delta H/RT_{n-1}} + e^{-\Delta H/RT_n}}{n} \right)} \right]$$

in which T_k is the mean kinetic temperature; ΔH is the heat of activation, 83.144kJ·mole^{-1}; R is the universal gas constant, 8.3144×10^{-3} kJ· mole^{-1}· degree^{-1}; T_1 is the value for the temperature (in degrees Kelvin [°K]) recorded during the first time period, T_2 is the value for the temperature recorded during the second time period, eg, second week; T_{n-1} is the value of the second to last time period, and T_n is the value for the temperature recorded during the nth time period. Typically, the time period is in days or weeks. The mean kinetic temperature determines the thermal exposure of a material. This allows an acceptable estimation to assess if a temperature excursion (or series of excursions) adversely affected a material.

FDA Guidelines provide recommendations for:

1. The design of stability studies to establish appropriate expiration dating periods and product storage requirements
2. The submission of stability information for investigational new drugs, biologicals, new drug applications, and biological product license applications

Thus, the guidelines represent a framework for the experimental design and data analysis as well as the type of documentation needed to meet regulatory requirements in the drug-development process.

Table 52-1. Stability Protocols

CONDITIONS	MINIMUM TIME PERIOD AT SUBMISSION
Long-term testing 25°C ± 2°C/60% ± 5% RH	12 mo
Accelerated testing 40°C ± 2°C/75% ± 5% RH	6 mo
Alternate testing[a] 30°C ± 2°C/65% ± 5% RH	12 mo

[a]Required if *significant change* occurs during 6-mo storage under conditions of accelerated testing.

Example Stability Pull Schedule for a Solid Oral Dose for Zone I and II

STORAGE CONDITIONS	DURATIONS (MONTHS)								
	0	1	3	6	9	12	18	24	36
25°C/60% RH	R*		X	X	X	X, Y	X	X	X
30°C/65% RH			O	O	O	O			
40°C/75% RH		X	X	X, Y					

*From Release testing if testing is within 30 days of stability set down.

R = Release Tests
- Appearance (visual)
- Identity
- Assay (HPLC)
- Impurities (HPLC)
- Dissolution (USP <711>)
- Moisture Content (Karl Fischer)

Uniformity of Dosage Unit

O = Pull and test only after 40°C/75% is out of specification
- Appearance (visual)
- Assay (HPLC)
- Impurities (HPLC)
- Dissolution (USP <711>)

X = Tests at Every Stability Pull
- Appearance (visual)
- Assay (HPLC)
- Impurities (HPLC)
- Dissolution (USP <711>)

Y = Additional tests periodically performed
- Moisture Content (Karl Fischer)

FDA Guidelines, however, has been reevaluated and revised significantly in the last few years, with the aim of harmonizing the technical requirements for the registration of pharmaceuticals worldwide. The International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) is a unique project that brought together regulatory authorities and experts from the pharmaceutical industry from three regions of the world; Europe, Japan, and the US. The first conference (ICH1) took place in November 1991 in Brussels, and the second conference (ICH2) in Orlando, FL, in October 1993. These conferences provided an open forum for discussion and resulted in the creation of an extensive set of guidelines dealing with the many aspects of safety, quality, and efficacy of medicinal products. The ICH Harmonized Tripartite Guideline provides a general indication on the requirements for *Stability Testing of New Drug Substances and Products.* The main thrust of the stability guideline centers around criteria for setting up stability protocols, shown in Table 52-1 and the example Stability Pull Schedule.

The guidelines were published in a draft form in the *Federal Register,* April 16, 1993. The final guidelines were published in 1994, with implementation of the guidelines occurring with Registration Applications after January 1, 1998. Revision 1 of the guidance was published in August 2001. Online computer can now access a complete listing of FDA publications and guidances. To view the publications, go to http://www.fda.gov/cder/guidance/index.htm.

PRODUCT STABILITY

Many factors affect the stability of a pharmaceutical product and include the stability of the active ingredient(s), the potential interaction between active and inactive ingredients, the manufacturing process, the dosage form, the container-liner-closure system, and the environmental conditions encountered during shipment, storage and handling, and length of time between manufacture and usage.

Classically, pharmaceutical product stability evaluations have been separated into studies of chemical (including biochemical) and physical stability of formulations. Realistically, there is no absolute division between these two arbitrary divisions. Physical factors, such as heat, light, and moisture, may initiate or accelerate chemical reactions, while every time a measurement is made on a chemical compound. Physical dimensions are included in the study.

In this treatment, physical and chemical stability are discussed along with those dosage form properties that can be measured and are useful in predicting shelf life. The effect of various physical and chemical phenomena of pharmaceuticals also is treated.

Knowledge of the physical stability of a formulation is very important for three primary reasons. First, a pharmaceutical product must appear fresh, elegant, and professional, for as long as it remains on the shelf. Any changes in physical appearance such as color fading or haziness can cause the patient or consumer to lose confidence in the product. Second, since some products are dispensed in multiple-dose containers, uniformity of dose content of the active ingredient over time must be ensured. A cloudy solution or a broken emulsion can lead to a non-uniform dosage pattern. Third, the active ingredient must be available to the patient throughout the expected shelf life of the preparation. A breakdown in the physical system can lead to non-availability or "dose dumping" of the medication to the patient. In the case of metered-dose inhaler pulmonary aerosols, particle aggregation may result in inadequate lung deposition of the medication.

The chemical causes of drug deterioration have been classified as incompatibility, oxidation, reduction, hydrolysis, racemization, and other mechanisms. In the latter category, decarboxylation, deterioration of hydrogen peroxide and hypochlorites, and the formation of precipitates have been included.

PHARMACEUTICAL DOSAGE FORMS

As the various pharmaceutical dosage forms present unique stability problems, they are discussed separately in the following section.

TABLETS—Stable tablets retain their original size, shape, weight, and color under normal handling and storage conditions throughout their shelf life. In addition, the *in vitro* availability of the active ingredients should not change appreciably with time.

Excessive powder or solid particles at the bottom of the container, cracks or chips on the face of a tablet, or appearance of crystals on the surface of tablets or on container walls are indications of physical instability of uncoated tablets. Hence, the effect of mild, uniform, and reproducible shaking and tumbling of tablets should be studied. The recommended test for such studies is the determination of tablet friability as described in the USP. Tablet Friability <1216> describes the recommended apparatus and the test procedure. After visual observation of the tablets for chips, cracks, and splits, the intact tablets are sorted and weighed to determine the amount of material worn away by abrasion. In general a maximum weight loss of not more than 1% of the weight of the tablets being tested is considered acceptable for most products. The results of these tests are comparative rather than absolute and should be correlated with actual stress experience. Packaged tablets also should be subjected to cross-country shipping tests as well as to various *drop tests.*

Tablet hardness (or resistance to crushing or fracturing) can be assessed by commercially available hardness testers. As results will vary with the specific make of the test apparatus used, direct comparison of results obtained on different instruments may not necessarily be made. Thus, the same instrument should be used consistently throughout a particular study.

Color stability of tablets can be followed by an appropriate colorimeter or reflectometer with heat, sunlight, and intense artificial light employed to accelerate the color deterioration. Caution must be used in interpreting the elevated temperature data, as the mechanism for degradation at that temperature may differ from that at a lower temperature. It is not always proper to assume that the same changes will occur at elevated temperatures as will be evidenced later at room temperature. Cracks, mottling, or tackiness of the coating indicates evidence of instability of coated tablets.

For tablets containing the more insoluble active ingredients, the results of dissolution tests are more meaningful than disintegration results for making bioavailability predictions. Dissolution-rate tests should be run in an appropriate medium such as artificial gastric and/or intestinal fluid at 37°. When no significant change (such as a change in the polymorphic form of the crystal) has occurred, an unaltered dissolution-rate profile of a tablet formulation usually indicates constant *in vivo* availability.

Uniformity of weight, odor, texture, drug and moisture contents, and humidity effect also are studied during a tablet stability test.

GELATIN CAPSULES—Hard gelatin capsules are the type used by pharmaceutical manufacturers in the production of the majority of their capsule products. The pharmacist in the extemporaneous compounding of prescriptions may also use hard gelatin capsules. Soft gelatin capsules are prepared from shells of gelatin to which glycerin or a polyhydric alcohol such as sorbitol has been added to render the gelatin elastic or plastic-like. Gelatin is stable in air when dry but is subject to microbial decomposition when it becomes moist or when it is maintained in aqueous solution. Normally hard gelatin capsules contain between 13% and 16% moisture. If stored in a high humidity environment capsule shells may soften, stick together, or become distorted and lose their shape. On the other hand, in an environment of extreme dryness gelatin capsules may harden and crack under slight pressure. Gelatin capsules should be protected from sources of microbial contamination.

Encapsulated products, like all other dosage forms, must be packaged properly.

Because moisture may be absorbed or released by gelatin capsules depending on the environmental conditions, capsules offer little physical protection to hygroscopic or deliquescent materials enclosed within a capsule when stored in an area of high humidity. It is not uncommon to find capsules packaged in containers along with a packet of desiccant material as a precautionary measure.

Both hard and soft gelatin capsules exposed to excessive heat and moisture may exhibit delayed or incomplete dissolution due to cross-linking of the gelatin in the capsule shell. The cross-linking of gelatin capsules is an irreversible chemical reaction. Cross-linking may also occur in capsules that are exposed to aldehydes and peroxides. Although cross-linked capsules may fail dissolution due to pellicle formation, digestive enzymes will dissolve the capsules. For hard or soft gelatin capsules that do not conform to the dissolution specification, the dissolution test may be repeated with the addition of enzymes. Where water or a medium with a pH less than 6.8 is specified as the medium in the individual monograph, the same medium specified may be used with the addition of purified pepsin that results in an activity of 750,000 units or less per 1000 mL. For media with a pH of 6.8 or greater, pancreatin can be added to produce not more than 1750 USP units of protease activity per 1000 mL.

SUSPENSIONS—A stable suspension can be redispersed homogeneously with moderate shaking and can be poured easily throughout its shelf life, with neither the particle-size distribution, the crystal form, nor the physiological availability of the suspended active ingredient changing appreciably with time.

Most stable pharmaceutical suspensions are flocculated; that is, the suspended particles are bonded together physically to form a loose, semi rigid structure. The particles are said to uphold each other while exerting no significant force on the liquid. Sedimented particles of a flocculated suspension can be redispersed easily at any time with only moderate shaking.

In nonflocculated suspensions, the particles remain as individuals unaffected by neighboring particles and are affected only by the suspension vehicle. These particles, which are smaller and lighter, settle slowly, but once they have settled, often form a hard, difficult-to-disperse sediment. Nonflocculated suspensions can be made acceptable by decreasing the particle size of the suspended material or by increasing the density and viscosity of the vehicle, thus reducing the possibility of settling.

When studying the stability of a suspension, first determine with a differential manometer if the suspension is flocculated. If the suspension is flocculated, the liquid will travel the same distance in the two side arms. With nonflocculated suspensions, the hydrostatic pressures in the two arms are unequal; hence, the liquids will be at different levels.

The history of settling of the particles of a suspension may be followed by a Brookfield viscometer fitted with a Helipath attachment. This instrument consists of a rotating T-bar spindle that descends slowly into the suspension as it rotates. The dial reading on the viscometer is a measure of the resistance that the spindle encounters at various levels of the sedimented suspension. This test must be run only on fresh, undisturbed samples.

An electronic particle counter and sizer, such as a Coulter counter, or a microscope may be used to determine changes in particle-size distribution. Crystal form alterations may be detected by microscopic, near-IR or Raman examination and, when suspected, must be confirmed by x-ray powder diffraction.

All suspensions should be subjected to cycling temperature conditions to determine the tendency for crystal growth to occur within the suspension. Shipping tests, ie, transporting bottles across the country by rail or truck are also used to study the stability of suspensions.

SOLUTIONS—A stable solution retains its original clarity, color, and odor throughout its shelf life. Retention of clarity of a solution is a main concern of a physical stability program. As visual observation alone under ordinary light is a poor test of clarity, a microscope light should be projected through a diaphragm into the solution. Undissolved particles will scatter the light, and the solution will appear hazy. While the Coulter counter also can be used, light-scattering instruments are the most sensitive means of following solution clarity.

Solutions should remain clear over a relatively wide temperature range such as 4 to 47°C. At the lower range an ingredient may precipitate due to its lower solubility at that temperature, while at the higher temperature the flaking of particles from the glass containers or rubber closures may destroy homogeneity. Thus, solutions should be subjected to cycling temperature conditions.

The stability program for solutions also should include a study of pH changes, especially when the active ingredients are soluble salts of insoluble acids or bases. Among other tests are observations for changes in odor, appearance, color, taste, light-stability, redispersibility, suspendibility, pourability, viscosity, isotonicity, gas evolution, microbial stability, specific gravity, surface tension, and pyrogen content, in the case of parenteral products.

When solutions are filtered, the filter medium may absorb some of the ingredients from the solution. Thus, the same type of filter should be used for preparing the stability samples as will be used to prepare the production-size batches.

For dry-packaged formulations reconstituted prior to use, the visual appearance should be observed on both the original dry material and on the reconstituted preparation. The color and odor of the cake, the color and odor of the solution, the moisture content of the cake, and the rate of reconstitution should be followed as a part of its stability profile.

EMULSIONS—A stable emulsion can be redispersed homogeneously to its original state with moderate shaking and can be poured at any stage of its shelf life. Although most of the important pharmaceutical emulsions are of the oil in water (O/W) type, many stability test methods can be applied to either an O/W or water in oil (W/O) emulsion.

Two simple tests are used to screen emulsion formulations. First, heating to 50 to 70°C and observing its gross physical stability either visually or by turbidimetric measurements can determine the stability of an emulsion. Usually the emulsion that is the most stable to heat is the one most stable at room temperature. However, this may not be true always, because an emulsion at 60°C may not be the same as it is at room temperature. Second, the stability of the emulsion can be estimated by the *coalescence time* test. Although this is only a rough quantitative test, it is useful for detecting gross differences in emulsion stability at room temperature.

Emulsions also should be subjected to refrigeration temperatures. An emulsion stable at room temperature has been found to be unstable at 4°C. It was reasoned that an oil-soluble emulsifier precipitated at the lower temperature and disrupted the system. An emulsion chilled to the extent that the aqueous base crystallizes is damaged irreversibly.

The ultracentrifuge also is used to determine emulsion stability. When the amount of separated oil is plotted against the time of centrifugation, a plateau curve is obtained. A linear graph results when the oil flotation (creaming) rate is plotted versus the square of the number of centrifuge revolutions per minute. The flotation rate is represented by the slope of the line resulting when the log distance of emulsion-water boundary from the rotor center is plotted against time for each revolution per minute.

For stability studies, two batches of an emulsion should be made at one time on two different sizes of equipment. One should be a bench-size lot and the other a larger, preferably production-size, batch. Different types of homogenizers produce different results, and different sizes of the same kind of homogenizer can yield emulsions with different characteristics.

OINTMENTS—Ointments have been defined as high-viscosity suspensions of active ingredients in a non-reacting

vehicle. A stable ointment is one that retains its homogeneity throughout its shelf-life period. The main stability problems observed in ointments are *bleeding* and changes in consistency due to aging or changes in temperature. When fluid components such as mineral oil separate at the top of an ointment, the phenomenon is known as *bleeding* and can be observed visually. Unfortunately, as there is no known way to accelerate this event, the tendency to *bleed* cannot be predicted.

An ointment that is too soft is messy to use, while one that is very stiff is difficult to extrude and apply. Hence, it is important to be able to define quantitatively the consistency of an ointment. This may be done with a penetrometer, an apparatus that allows a pointed weight to penetrate into the sample under a measurable force. The depth of the penetration is a measure of the consistency of an ointment. Consistency also can be measured by the Helipath attachment to a high-viscosity viscometer or by a Burrell Severs rheometer. In the latter instrument, the ointment is loaded into a cylinder and extruded with a measured force. The amount extruded is a measure of the consistency of the ointment.

Ointments have a considerable degree of structure that requires a minimum of 48 hours to develop after preparation. As rheological data on a freshly made ointment may be erroneous, such tests should be performed only after the ointment has achieved equilibrium. Slight changes in temperature (1 or 2°C) can affect the consistency of an ointment greatly; hence, rheological studies on ointments must be performed only at constant and controlled temperatures.

Among the other tests performed during the stability study of an ointment are a check of visual appearance, color, odor, viscosity, softening range, consistency, homogeneity, particle-size distribution, and sterility. Undissolved components of an ointment may change in crystal form or in size with time. Microscopic examination or an x-ray diffraction measurement may be used to monitor these parameters.

In some instances it is necessary to use an ointment base that is less than ideal, to achieve the required stability. For example, drugs that hydrolyze rapidly are more stable in a hydrocarbon base than in a base containing water, even though they may be more effective in the latter.

TRANSDERMAL PATCHES—A typical transdermal patch consists of a protective backing, a matrix containing active drug, an adhesive that allows the patch to adhere to the skin, and a release liner to protect the skin adhering adhesive. Therefore, the transdermal patch must deliver drug as labeled, adhere properly to both the backing and to the patient's skin. In addition, the transdermal patch must be pharmaceutically elegant through the shelf life of the product. For a transdermal patch, this means that the release line peels easily with minimal transfer of adhesive onto the release liner and that the adhesive does not ooze from the sides of the patch. Therefore, the typical stability related tests for transdermal patches are, appearance, assay, impurities, drug release USP<724> and, backing peel force.

METERED-DOSE AEROSOLS DRUG PRODUCTS—A metered dose inhalation product consists of an aerosol can containing a propellant, a drug and a mouthpiece used to present an aerosolized drug to the patient. There are many drug contact components in a metered-dose inhalation product. Therefore, the drug may be in contact with materials that could allow plasticizer leach into the drug. The typical stability related tests for metered-dose aerosols include appearance, assay, impurities, plume geometry, emitted dose, particle size distribution of the emitted dose, and number of doses per unit. In addition, stability studies on leachables may be required. Shelf life of metered-dose aerosols drug products may also be dependent on the orientation that the drug product is stored. Typically most canisters type product are tested at least in the upright orientation.

DRY-POWDERED INHALATION PRODUCTS—A dry-powdered inhalation product consists of drug with excipients delivered in a dry powdered form. The delivery system for a dry-powdered inhalation product may be a separate device or integrated with the active. A dry-powdered dosage must reproducibly deliver a specific amount of drug at a particle size that can be deposited into the lungs. Particles too large will get trapped in the throats and particles too small will just be carried out of the lungs on the next expiration. The typical stability related tests for dry powder inhalation products include appearance, assay, impurities, emitted dose, particle size distribution of the emitted dose, and water content.

NASAL INHALATION PRODUCTS—A nasal inhalation product consists of drug with excipients delivered from a delivery system. The delivery system for a nasal inhalation product may be a separate device or integrated with the active. A nasal inhalation product must reproducibly deliver a specific amount of drug at a particle size and plume that can be deposited into the nasal membrane. Particles too large will not be absorbed into nasal membrane or run out of the nose; and poor spray pattern will deposit the drug ineffective in the nasal cavity. The typical stability related tests for nasal inhalation products include appearance, assay, impurities, spray content uniformity, particle (droplet) size distribution of the emitted dose, spray pattern or /and plume geometry, leachables, weight loss and preservative content. Sterility and microbial testing may be required periodically for stability testing.

INCOMPATIBILITY

Typically, physicochemical stability is assessed at the preformulation stage of development. A drug substance candidate is treated with acid, base, heat, light, and oxidative conditions to assess its inherit chemical stability. Binary mixtures of the drug substance with individual excipients are also investigated at the preformulation stage. These tests are performed to determine the drug substance sensitivity to degrade or react with common pharmaceutical excipients. The most common reactions observed for drug substances from these tests include: hydrolysis, epimerization (racemization), decarboxylation, dehydration, oxidation, polymerization, photochemical decomposition, and addition. All drug substances have the potential to degrade by at least one of the reactions mentioned above. With an understanding of the stability/reactivity of a drug substance in the preformulation stage, it is possible to formulate the drug product to minimize drug decomposition. Numerous examples are described in other sections of this book, and the literature is replete with illustrations.

While undesirable reactions between two or more drugs are said to result in a *physical, chemical,* or *therapeutic* incompatibility, physical incompatibility is somewhat of a misnomer. It has been defined as a physical or chemical interaction between two or more ingredients that leads to a visibly recognizable change. The latter may be in the form of a gross precipitate, haze, or color change.

On the other hand, a chemical incompatibility is classified as a reaction in which a visible change is not necessarily observed. Since there is no visible evidence of deterioration, this type of incompatibility requires trained, knowledgeable personnel to recognize it.

A therapeutic incompatibility has been defined as an undesirable pharmacological interaction between two or more ingredients that leads to

1. Potentiation of the therapeutic effects of the ingredients
2. Destruction of the effectiveness of one or more of the ingredients
3. Occurrence of a toxic manifestation within the patient.

REACTION KINETICS

An understanding of reaction kinetics is important in determining the shelf life of a product.

CHEMICAL REACTIONS

The most frequently encountered chemical reactions, which may occur within a pharmaceutical product, are described below.

OXIDATION-REDUCTION—Oxidation is a prime cause of product instability, and often, but not always, the addition of oxygen or the removal of hydrogen is involved. When molecular oxygen is involved, the reaction is known as auto-oxidation because it occurs spontaneously, though slowly, at room temperature.

Oxidation, or the loss of electrons from an atom, frequently involves free radicals and subsequent chain reactions. Only a very small amount of oxygen is required to initiate a chain reaction. In practice, it is easy to remove most of the oxygen from a container, but very difficult to remove it all. Hence, nitrogen and carbon dioxide frequently are used to displace the headspace air in pharmaceutical containers to help minimize deterioration by oxidation.

As an oxidation reaction is complicated, it is difficult to perform a kinetic study on oxidative processes within a general stability program. The redox potential, which is constant and relatively easy to determine, can, however, provide valuable predictive information. In many oxidative reactions, the rate is proportional to the concentration of the oxidizing species but may be independent of the concentration of the oxygen present. The rate is influenced by temperature, radiation, and the presence of a catalyst. An increase in temperature leads to an acceleration in the rate of oxidation. If the storage temperature of a preparation can be reduced to 0 to 5°C, usually it can be assumed that the rate of oxidation will be at least halved.

The molecular structures most likely to oxidize are those with a hydroxyl group directly bonded to an aromatic ring (eg, phenol derivatives such as catecholamines and morphine), conjugated dienes (eg, vitamin A and unsaturated free fatty acids), heterocyclic aromatic rings, nitroso and nitrite derivatives, and aldehydes (eg, flavorings). Products of oxidation usually lack therapeutic activity. Visual identification of oxidation, for example, the change from colorless epinephrine to its amber colored products, may not be visible in some dilutions or to some eyes.

Oxidation is catalyzed by pH values that are higher than optimum, polyvalent heavy metal ions (eg, copper and iron), and exposure to oxygen and UV illumination. The latter two causes of oxidation justify the use of antioxidant chemicals, nitrogen atmospheres during ampul and vial filling, opaque external packaging, and transparent amber glass or plastic containers.

Trace amounts of heavy metals such as cupric, chromic, ferrous, or ferric ions may catalyze oxidation reactions. As little as 0.2 mg of copper ion per liter considerably reduces the stability of penicillin. Similar examples include the deterioration of epinephrine, phenylephrine, lincomycin, isoprenaline, and procaine hydrochloride. Adding chelating agents to water to sequester heavy metals and working in special manufacturing equipment (eg, glass) are some means used to reduce the influence of heavy metals on a formulation. Parenteral formulations should not come in contact with heavy metal ions during their manufacture, packaging, or storage.

Hydronium and hydroxyl ions catalyze oxidative reactions. The rate of decomposition for epinephrine, for example, is more rapid in a neutral or alkaline solution with maximum stability (minimum oxidative decomposition) at pH 3.4. There is a pH range for maximum stability for any antibiotic and vitamin preparation, which usually can be achieved by adding an acid, alkali, or buffer.

Oxidation may be inhibited by the use of antioxidants, called negative catalysts. They are very effective in stabilizing pharmaceutical products undergoing a free-radical-mediated chain reaction. These substances, which are easily oxidizable, act by possessing lower oxidation potentials than the active ingredient. Thus, they undergo preferential degradation or act as chain inhibitors of free radicals by providing an electron and receiving the excess energy possessed by the activated molecule.

The ideal antioxidant should be stable and effective over a wide pH range, soluble in its oxidized form, colorless, nontoxic, nonvolatile, nonirritating, effective in low concentrations, thermostable, and compatible with the container-closure system and formulation ingredients.

The commonly used antioxidants for aqueous systems include sodium sulfite, sodium metabisulfite, sodium bisulfite, sodium thiosulfate, and ascorbic acid. For oil systems, ascorbyl palmitate, hydroquinone, propyl gallate, nordihydroguaiaretic acid, butylated hydroxytoluene, butylated hydroxyanisole, and alpha-tocopherol are employed.

Synergists, which increase the activity of antioxidants, are generally organic compounds that complex small amounts of heavy metal ions. These include the ethylenediamine tetraacetic acid (EDTA) derivatives, dihydroethylglycine, and citric, tartaric, gluconic, and saccharic acids. EDTA has been used to stabilize ascorbic acid, oxytetracycline, penicillin, epinephrine, and prednisolone.

Reduction reactions are much less common than oxidative processes in pharmaceutical practice. Examples include the reduction of gold, silver, or mercury salts by light to form the corresponding free metal.

HYDROLYSIS—Drugs containing esters (eg, cocaine, physostigmine, aspirin, tetracaine, procaine and methyldopa), amides (eg, dibucaine), imides (eg, amobarbital), imines (eg, diazepam) and lactam (eg, penicillins, cephalosporins) functional groups are among those prone to hydrolysis.

Hydrolysis reactions are often pH dependent and are catalyzed by either hydronium ion or hydroxide ions (specific-acid or specific-base catalysis, respectively). Hydrolysis reactions can also be catalyzed by either a Brønsted acid or a Brønsted base (general-acid or general-base catalysis, respectively). Sources of Brønsted acid or base include buffers and some excipients. Sometimes, it is necessary to compromise between the optimum pH for stability and that for pharmacological activity. For example, several local anesthetics are most stable at a distinctly acid pH, whereas for maximum activity they should be neutral or slightly alkaline. Small amounts of acids, alkalines, or buffers are used to adjust the pH of a formulation. Buffers are used when small changes in pH are likely to cause major degradation of the active ingredient.

Obviously, the amount of water present can have a profound effect on the rate of a hydrolysis reaction. When the reaction takes place fairly rapidly in water, other solvents sometimes can be substituted. For example, barbiturates are much more stable at room temperature in propylene glycol–water than in water alone.

Modification of chemical structure may be used to retard hydrolysis. In general, as it is only the fraction of the drug in solution that hydrolyzes, a compound may be stabilized by reducing its solubility. This can be done by adding various substituents to the alkyl or acyl chain of aliphatic or aromatic esters or to the ring of an aromatic ester. In some cases less-soluble salts or esters of the parent compound have been found to aid product stability. Steric and polar complexation have also been employed to alter the rate of hydrolysis. Caffeine reduces the rate of hydrolysis and thus promotes stability by complexation with local anesthetics such as benzocaine, procaine, or tetracaine.

Esters and β-lactams are the chemical bonds that are most likely to hydrolyze in the presence of water. For example, the acetyl ester in aspirin is hydrolyzed to acetic acid and salicylic acid in the presence of moisture, but in a dry environment the hydrolysis of aspirin is negligible. The aspirin hydrolysis rate increases in direct proportion to the water vapor pressure in an environment.

The amide bond also hydrolyzes, though generally at a slower rate than comparable esters. For example, procaine (an ester) will hydrolyze upon autoclaving, but procainamide will not. The amide or peptide bond in peptides and proteins varies

in the labiality to hydrolysis. The lactam and azomethine (or imine) bonds in benzodiazepines are also labile to hydrolysis. The major chemical accelerators or catalysts of hydrolysis are adverse pH and specific chemicals (eg, dextrose and copper in the case of ampicillin hydrolysis).

The rate of hydrolysis depends on the temperature and the pH of the solution. A much-quoted estimation is that for each 10°C rise in storage temperature, the rate of reaction doubles or triples. As this is an empiricism, it is not always applicable.

When hydrolysis occurs, the concentration of the active ingredient decreases while the concentration of the decomposition products increases. The effect of this change on the rate of the reaction depends on the order of the reaction. With zero-order reactions the rate of decomposition is independent of concentration of the ingredient. Although dilute solutions decompose at the same absolute rate as more concentrated solutions, the more dilute the solution, the greater the proportion of active ingredient destroyed in a given time; ie, the percentage of decomposition is greater in more dilute solutions. Increasing the concentration of an active ingredient that is hydrolyzing by zero-order kinetics will slow the percentage decomposition.

With first-order reactions, which occur frequently in the hydrolysis of drugs, the rate of change is directly proportional to the concentration of the reactive substance. Thus, changes in the concentration of the active ingredient have no influence on the percentage decomposition.

The degradation of many drugs in solution accelerates or decelerates exponentially as the pH is decreased or increased over a specific range of pH values. Improper pH ranks with exposure to elevated temperature as a factor most likely to cause a clinically significant loss of drug, resulting from hydrolysis and oxidation reactions. A drug solution or suspension, for example, may be stable for days, weeks, or even years in its original formulation, but when mixed with another liquid that changes the pH, it degrades in minutes or days. It is possible that a pH change of only one unit (eg, from 4 to 3 or 8 to 9) could decrease drug stability by a factor of ten or greater.

A pH-buffer system, which is usually a weak acid or base and its salt, is a common excipient used in liquid preparations to maintain the pH in a range that minimizes the drug degradation rate. The pH of drug solutions may also be either buffered or adjusted to achieve drug solubility. For example, pH in relation to pKa controls the fractions of the usually more soluble ionized and less soluble nonionized species of weak organic electrolytes.

INTERIONIC (ION N+ −ION N−) COMPATIBILITY— The compatibility or solubility of oppositely charged ions depends mainly on the number of charges per ion and the molecular size of the ions. In general, polyvalent ions of opposite charge are more likely to be incompatible. Thus, an incompatibility is likely to occur upon the addition of a large ion with a charge opposite to that of the drug.

As many hydrolytic reactions are catalyzed by both hydronium and hydroxyl ions, pH is an important factor in determining the rate of a reaction. The pH range of minimum decomposition (or maximum stability) depends on the ion having the greatest effect on the reaction. If the minimum occurs at about pH 7, the two ions are of equal effect. A shift of the minimum toward the acid side indicates that the hydroxyl ion has the stronger catalytic effect and *vice versa* in the case of a shift toward the alkaline side. In general, hydroxyl ions have the stronger effect. Thus, the minimum is often found between pH 3 and 4. The influence of pH on the physical stability of two-phase systems, especially emulsions, is also important. For example, intravenous fat emulsion is destabilized by acidic pH.

DECARBOXYLATION—Pyrolytic solid-state degradation through decarboxylation usually is not encountered in pharmacy, as relatively high heats of activation (25 to 30 kcal) are required for the reaction. However, solid *p*-aminosalicylic acid undergoes pyrolytic degradation to *m*-aminophenol and carbon dioxide. The reaction, which follows first-order kinetics, is highly pH-dependent and is catalyzed by hydronium ions. The decarboxylation of *p*-aminobenzoic acid occurs only at extremely low pH values and at high temperatures.

Some dissolved carboxylic acids, such as *p*-aminosalicylic acid, lose carbon dioxide from the carboxyl group when heated. The resulting product has reduced pharmacological potency. β-Keto decarboxylation can occur in some solid antibiotics that have a carbonyl group on the β-carbon of a carboxylic acid or a carboxylate anion. Such decarboxylations will occur in the following antibiotics: carbenicillin sodium, carbenicillin free acid, ticarcillin sodium, and ticarcillin free acid.

RACEMIZATION—Racemization, or the action or process of changing from an optically active compound into a racemic compound or an optically inactive mixture of corresponding *R* (*rectus*) and *S* (*sinister*) forms, is a major consideration in pharmaceutical stability. Optical activity of a compound may be monitored by polarimetry and reported in terms of specific rotation. Chiral HPLC has been used in addition to polarimetry to confirm the enantiomeric purity of a sample.

In general, racemization follows first-order kinetics and depends on temperature, solvent, catalyst, and the presence or absence of light. Racemization appears to depend on the functional group bound to the asymmetric carbon atom, with aromatic groups tending to accelerate the process.

EPIMERIZATION—Members of the tetracycline family are most likely to incur epimerization. This reaction occurs rapidly when the dissolved drug is exposed to a pH of an intermediate range (higher than 3), and it results in the steric rearrangement of the dimethylamino group. The epimer of tetracycline, epitetracycline, has little or no antibacterial activity.

PHOTOCHEMICAL REACTIONS

Photolytic degradation can be an important limiting factor in the stability of pharmaceuticals. A drug can be affected chemically by radiation of a particular wavelength only if it absorbs radiation at that wavelength and the energy exceeds a threshold. Ultraviolet radiation, which has a high energy level, is the cause of many degradation reactions. Exposure to, primarily, UV illumination may cause oxidation (photo-oxidation) and scission (photolysis) of covalent bonds. Nifedipine, nitroprusside, riboflavin, and phenothiazines are very labile to photo-oxidation. In susceptible compounds, photochemical energy creates free radical intermediates, which can perpetuate chain reactions.

If the absorbing molecule reacts, the reaction is said to be photochemical in nature. When the absorbing molecules do not participate directly in the reaction, but pass their energy to other reacting molecules, the absorbing substance is said to be a photosensitizer.

As many variables may be involved in a photochemical reaction, the kinetics can be quite complex. The intensity and wavelength of the light and the size, shape, composition, and color of the container may affect the velocity of the reaction.

The photodegradation of chlorpromazine through a semiquinone free-radical intermediate follows zero-order kinetics. On the other hand, alcoholic solutions of hydrocortisone, prednisolone, and methylprednisolone degrade by reactions following first-order kinetics.

Colored-glass containers most commonly are used to protect light-sensitive formulations. Yellow-green glass gives the best protection in the ultraviolet region, while amber confers considerable protection from ultraviolet radiation but little from infrared. Riboflavin is best protected by a stabilizer that has a hydroxyl group attached to or near the aromatic ring. The photodegradation of sulfacetamide solutions may be inhibited by an antioxidant such as sodium thiosulfate or metabisulfite.

A systematic approach to photostability testing is recommended covering, as appropriate, studies such as tests on the drug substance, tests on the exposed drug product outside of the immediate pack; and if necessary, tests on the drug product

in the immediate pack. ICH Q1B discusses the minimum requirements for assessing photostability. Drug substance is first assessed by exposing sample powder having a depth of not more than 3 mm to an overall illumination of not less than 1.2 million lux hours and an integrated near ultraviolet energy of not less than 200 watt hours/square meter. If the drug substance shows sensitivity to photodegrations, then the drug product will need to be tested as well. The testing of drug product uses the same light exposure that was used to test drug substance. The drug product should be tested directly exposed to light and in its container closure system.

ULTRASONIC ENERGY

Ultrasonic energy, which consists of vibrations and waves with frequencies greater than 20,000 Hz, promotes the formation of free radicals and alters drug molecules. Changes in prednisolone, prednisone acetate, or deoxycorticosterone acetate suspensions in an ultrasonic field have been observed spectrometrically in the side chain at C-17 and in the oxo group of the A ring. With sodium alginate, in an ultrasonic field, it has been reported that above a minimum power output, degradation increased linearly with increased power.

IONIZING RADIATION

Ionizing radiation, particularly gamma rays, has been used for the sterilization of certain pharmaceutical products. At the usual sterilizing dose, 2.5 mRad, it seldom causes appreciable chemical degradation. In general, formulations that are in the solid or frozen state are more resistant to degradation from ionizing radiation than those in liquid form. For example, many of the vitamins are little affected by irradiation in the solid state but are decomposed appreciably in solution. On the other hand, both the liquid- and solid-state forms of atropine sulfate are affected seriously by radiation.

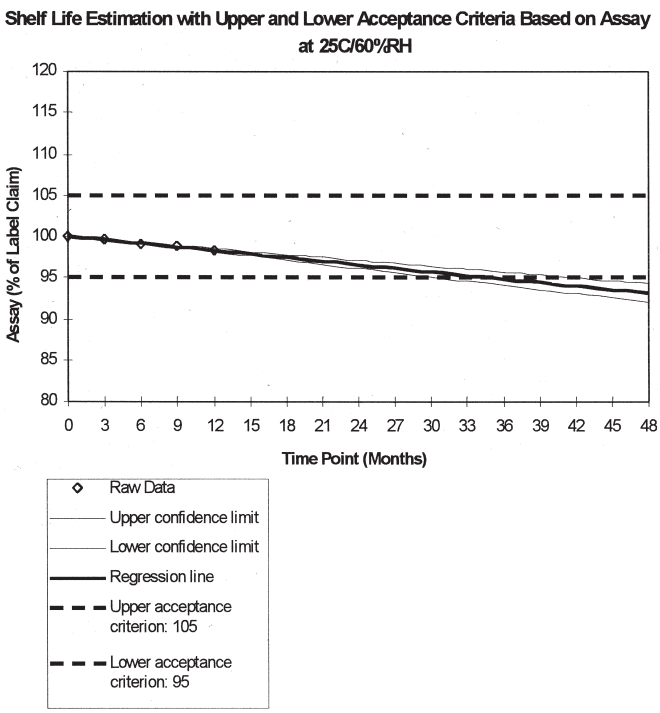


Figure 52-1. Typical two-sided shelf-life estimation plot.

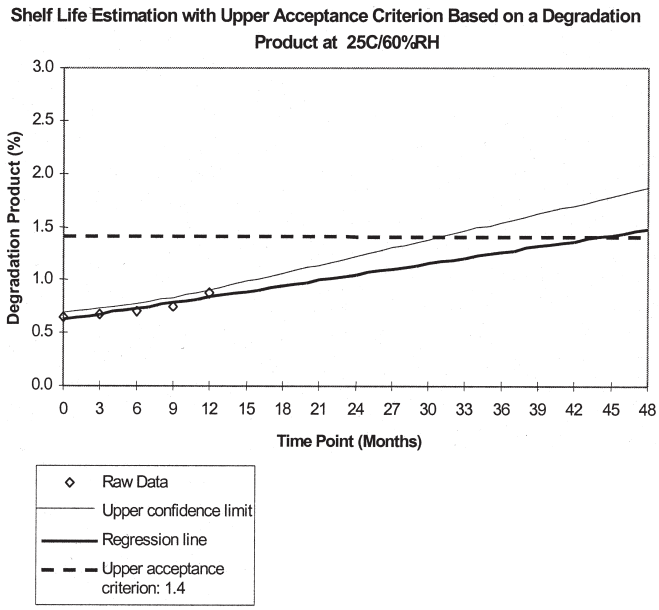


Figure 52-2. Typical one-sided shelf-life estimation plot.

PREDICTING SHELF LIFE

ICH Recommended Evaluation

The shelf life of a commercial drug product must be determined in the commercial container closure at the defined storage conditions. ICH requires at least 12 months stability data at the time of NDA submission. Most products require at least 24 months to be commercially viable. The ICH Q1E recommends how the 12 months data may be used to predict long-term stability. Figures 52-1 and 52-2 show trending graphs with double-sided and single-sided 95% confidence limits plots, respectively.

Figure 52-1 shows a plot of 12 months of assay (potency) results versus time. The acceptance criteria for this test have a lower and an upper limit of 95% and 105%, respectively. The extrapolated line from this data set intersects the lower acceptance limit at about 35 months. However, there is always statistical uncertainty when extrapolating a data set. The 95% confidence limit is used to take this uncertainty into account. The lower 95% confidence intersects the lower acceptance limit at about 29 months. Therefore, this product would be assessed an expiration date of 29 months.

Figure 52-2 shows a plot of 12 months of degradation product results. In this case, the acceptance criterion is an upper limit of not more than 1.4%. The extrapolated line from this data set intersects the acceptance limit at about 44 months. The upper 95% confidence limit curve intersects the acceptance limit at about 30 months. Therefore, this product would be assessed an expiration date of 30 months. The expiration of a product is the time where the confidence line intersects with the acceptance limit. Trend analysis of data need only be performed on test data that shows a change related to time.

Approximations in Assessing Product Stability—Estimation of Temperature Effect

In early development, a shelf life prediction of a clinical material, especially a Phase I material, may be based on a very limited amount of sample and limited amount of time to make the evaluation. One way to estimate long-term storage for a material is by extrapolating data from studies performed at elevated conditions. An understanding of potential activation energy is needed to estimate long-term stability. Many may have heard

of the estimate that for every 10°C decrease in storage temperature the shelf-life doubles. This is only true, however, if the activation energy of the reaction(s) that causes degradation is 15 kcal/moles. The activation energy, E_a, for many chemical processes related to the degradation of a drug substance/product is typically within the range of 10 to 25 kcal/moles.

The equation below shows a way of calculating the $Q_{\Delta T}$ value that may be used to estimate the affect of temperature on shelf life.

$$Q_{\Delta T} = \exp\left[\frac{E_a}{R}\left(\frac{\Delta T}{T + \Delta T(T)}\right)\right] \quad (1)$$

where, $Q_{\Delta T}$ is a factor (multiplier/divisor) used to estimate the change in the reaction rate constant with change in temperature, ΔT. E_a is the activation energy established for a reaction

An approximation for the change in reaction rate constants due to the temperature effects are shown in the table below.

Ea (kcal/mole)	Q_5 (25 to 30°C)	Q_{10} (25 to 35°C)	Q_{15} (25 to 40°C)
10	1.32	1.73	2.24
15	1.52	2.27	3.36
20	1.75	2.99	5.04
25	2.01	3.93	7.55

Therefore, the old rule of thumb that a reaction rate doubles with every 10°C is only true if the reaction has an activation energy between 10 to 15 kcal/mole (Q_{10} = 1.73 and 2.27, respectively). Q_{15} is useful for understanding the relationship of ICH accelerated temperature of 40°C has with controlled room temperature at 25°C. Materials made and packaged for clinical studies are usually tested at an accelerated condition in order to predict that the packaged material will be stable for the duration of the clinical study. A material stable for one month at accelerated temperature (40°C) supports that the material stored at room temperature should be stable for at least 3 months. This true only when the activation energy of the degradation process is about 15 kcal/mole (Q_{15} factor = 3.36) [In other words, a reaction at 40°C should be 3.36 times faster than the same reaction at 25°C; or the reaction will take 3.36 times longer at 25°C than at 40°C)].

The technique of estimating the shelf life of a formulation from its accumulated stability data has evolved from examining the data and making an educated guess through plotting the time-temperature points on appropriate graph paper and crudely extrapolating a regression line to the application of rigorous physical-chemical laws, statistical concepts, and computers to obtain meaningful, reliable estimates.

A simple means of estimating shelf life from a set of computer-prepared tables has been described by Lintner et al.[6] This system was developed to select the best prototype formulation on the basis of short-term stability data and predict both estimated and minimum shelf-life values for the formulation. It is a middle-ground approach between the empirical methods and the modern, rigorous statistical concepts. All calculations can be made readily by hand, and the estimated values can be obtained easily from appropriate tables. The system assumes that

1. Shelf-life predictions can be made satisfactorily for lower temperatures using the classical Arrhenius model from data obtained at higher temperatures.
2. The energy of activation of the degradation reaction is between 10 and 20 kcal/mol (this is a safe assumption, as Kennon[8] has noted that rarely are drugs with energies of activation below 10 kcal/mol used in pharmacy, and for values as high as 20 kcal/mol, the error in the shelf-life prediction will be on the conservative side).
3. The rate of decomposition will not increase beyond that already observed.
4. The standard deviation of the replicated assays is known or can be estimated from the analytical data.

This concept further assumes that the degradation reaction follows zero- or pseudo-zero-order kinetics. For data corresponding to a zero-, first-, or second-order degradation pattern, it is impossible to distinguish one order from another with usual analytical procedures, when the total degraded material is not large. In addition, shelf-life calculations assuming zero-order kinetics are more conservative than those for higher orders.

This middle-ground system is useful in creating the experimental design for the stability study. The formulator has the opportunity to study various combinations of parameters to try to optimize the physical-statistical model. One can check the effect of improving the assay standard deviation, running additional replicates, using different time points, and assuming various degradation rates and energies of activation on the stability of the test formulation.

McMinn and Lintner later developed and reported on an information-processing system for handling product stability data.[9] This system saves the time of formulators in analyzing and interpreting their product stability data, in addition to minimizing the amount of clerical help needed to handle an ever-increasing assay load. For products such as those of vitamins, for example, where large overages are required, the statistical portions of this advanced technique aid the manufacturer to tailor the formula composition to obtain the desired and most economical expiration dating.

This system stores both physical and chemical data and retrieves the information in three different formats (one of which was designed specifically for submitting to regulatory agencies). It analyzes single-temperature data statistically by analysis of covariance and regression or multiple-temperature data by weighted or unweighted analysis using the Arrhenius relationship; provides estimates of the shelf life of the preparation with the appropriate confidence intervals; preprints the assay request cards that are used to record the results of the respective assay procedures and to enter the data into the system; and produces a 5-yr master-stability schedule as well as periodic 14-day schedules of upcoming assays.

As mentioned above, a portion of the advanced system analyzes the stability data obtained at a single temperature by analysis of covariance and regression. This analysis is based on the linear (zero-order) model

$$Y_{ij} = \beta_i X_{ij} + \alpha_i + \varepsilon_{ij} \quad (2)$$

where Y_{ij} is the percentage of label of the jth stability assay of the ith lot, X_{ij} is the time in months at which Y_{ij} was observed, β_i and α_i are the slope and intercept, respectively, of the regression line of the ith lot, and ε_{ij} is a random error associated with Y_{ij}. The random errors are assumed to be identically and independently distributed normal variables with a zero mean and a common variance, σ^2.

A summary of the regression analysis for each individual lot and for the combination of these lots, plus a summary of the analyses of covariance and deviation from regression is prepared by the computer.

Because the computer combines, or pools, the stability data from the individual lots, irrespective of the statistical integrity of this step, the pooled data are examined for validity by the F test. The mean square of the regression coefficient (slope) is divided by the mean square of the deviation within lots, and similarly, the adjusted mean (y intercept) is divided by the common mean square to give the respective F ratios. The latter values then are compared with the critical 5% F values. When the calculated F values are smaller than the critical F values, the data may be combined, and the pooled data analyzed.

A printout for the combined lots as well as for each individual lot provides the estimated rate of degradation and its standard error in percentage per month for each ingredient. The *Student t* value is calculated from these estimates and tested for significance from zero. When the t value is significant, the printout contains an estimate of the shelf life with the appropriate confidence interval. When the t value is not significantly different from zero, estimates of the minimum and projected shelf-life values are made. In addition, coordinates of the calculated least-squares regression line with ap-

propriate confidence limits for the mean and individual predicted assays are printed.

Plots of the resulting least-squares line containing the individual data points also are printed by the computer. For the calculation of X_0, $\hat{Y}$ equals $\overline{Y} + \hat{\beta}(X_0 - \overline{X}..)$, where $\hat{\beta}$ is the least-squares estimate of the slope, and $\overline{X}..$ is the mean time of assay.

The sample variance for this estimate, $S^2(\hat{Y})$ is equal to

$$S_{y\cdot x}{}^2\left[\frac{1}{N} + \left[\frac{(X_0 - \overline{X}..)^2}{\sum(X_{ij} - \overline{X}..)^2}\right]\right] \quad (3)$$

where N is the number of assays. The 95% confidence interval is equal to $Y \pm t_{0.05S(\hat{Y})}$.

For cases in which the slope of the best fitting line is positive and significantly different from zero (resulting, eg, from solvent evaporation), the statement "no degradation has been detected and hence no shelf-life estimate is made" is printed. When the computed line has a positive slope but not significantly different from zero, only the minimum shelf-life value is calculated.

Traditionally, extensive stability data are collected at the recommended storage temperatures (usually refrigerator and/or room temperature) to be placed on the label of the package. However, elevated-temperature data are very valuable in determining the shelf life of a product. In practice, multiple levels of thermal stress are applied to the formulation so that appropriate shelf-life estimates can be made for normally expected marketing conditions. In cases in which data from accelerated studies are used to project a tentative expiration date that is beyond the date supported by actual shelf-life studies, testing must continue until the tentative expiration date is verified.

The effect of temperature variation on the rate of a reaction can be expressed by an integrated form of the Arrhenius equation

$$k = se^{-E_A/\mathrm{RT}} \quad (4)$$

where, k is the rate constant, E_A is the energy of activation in kcal/mole, R is the universal gas constant of 1.987 cal/deg mole, T is the temperature in degrees in Kelvin, and S is a constant that is related to the specific reaction.

$$\log\frac{k_2}{k_1} = \frac{E_a}{2.303\mathrm{R}}\left(\frac{T_2 - T_1}{T_2 * T_1}\right) \quad (5)$$

where, k_1 is the rate constant at temperature T_1 and k_2 is the rate constant at temperature T_2.

A weighted modification of this model has been incorporated into the previously described computerized system. Each printout contains a statement concerning the acceptability of the Arrhenius assumption with its appropriate probability level, the slope and intercept for the Arrhenius line, the estimated apparent energy of activation with its 95% confidence limits, plus the estimated shelf-life values at selected temperatures.

The analysis of first-order stability data is based on the linear model

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + \varepsilon_{ij} \quad (6)$$

where Y_{ij} is the natural logarithm of the assay value for the jth observation of the ith temperature, X_{ij} is the elapsed time in months for the assay sample for the ith temperature, β_i and α_i are the slope and intercept, respectively, and ε_{ij} is a random error associated with Y_{ij}. The errors are assumed to be distributed identically and independently, normally with a zero mean and variance σ^2.

For orders other than first, Y_{ij} represents the concentration raised to the power of 1 minus the order.

The estimated rate constant (ie, the negative slope) is

$$-b_i = -\sum_j (Y_{ij} - Y_i)(X_{ij} - X_i)/\sum_j (X_{ij} - X_i)^2 \quad (7)$$

The standard error of the estimated rate constant is

$$S_{-b_i} = \frac{S(X/Y)}{[\sum(X_{ij} - X_i)^2]^{1/2}} \quad (8)$$

where $S(Y/X)$, the residual standard error, is equal to

$$S(X/Y) = \left\{\frac{1}{N-2}\left[\sum_{j=1}^{12}(Y_{ij} - Y_i)^2 - \frac{[\sum(X_{ij} - X_i)(Y_{ij} - Yi)^2]}{\sum(X_{ij} - X_i)^2}\right]\right\}^{1/2} \quad (9)$$

According to the Arrhenius relationship, faster degradation occurs at the higher temperatures; hence, assays for the high-temperature data usually are run more often but for a shorter period of time. The effect of simple least-squares analysis of this type of data is to force the Arrhenius equation through the low temperature data and essentially ignore the high-temperature information. Thus, much more credence is placed in the point estimates of the low temperature than is warranted. In addition, the usual confidence limits on extrapolated degradation rates at refrigerator or room temperature cannot be made validly. For these reasons, Bentley[10] presented a method based on a weighted least-squares analysis to replace the unweighted approximation. He also developed a statistical test for the validity of the Arrhenius assumption, which is computed easily from the results of the unweighted method.

To make shelf-life estimates from elevated temperature data, two storage temperatures are obviously the minimum. As the accuracy of the extrapolation is enhanced by using additional temperatures, a minimum of four different temperatures is recommended for most product stability studies. With the current use of computers to do the bulk of stability calculations, including weighted least-squares analysis, the temperatures and storage conditions need not be selected for arithmetic convenience.

It is not necessary to determine the mechanism of the degradation reaction. In most cases, it is necessary only to follow some property of degradation and to linearize this function. Either the amount of intact drug or the amount of a formed degradation product may be followed. It usually is impractical to determine the exact order of the reaction. With assay errors in the range of 2 to 5%, at least 50% decomposition must occur before the reaction order can be determined. As the loss with pharmaceuticals generally is less, zero-order kinetics should be assumed, unless the reaction order is known from previous work. In any case, replication of stability assays is advisable.

The batches of drugs used for a stability study should be representative of production run material or at least material of a known degree of purity. The quality of the excipients also should be known, as their impurities or even their moisture content can affect product stability deleteriously. Likewise, the samples of the formulation taken for the stability study must be representative of the lot.

Specific, stability-indicating assay methods must be used, to make meaningful shelf-life estimates. The reliability and specificity of the test method on the intact molecule and on the degradation products must be demonstrated.

ADDITION OF OVERAGE

The problem of declining potency in an unstable preparation can be ameliorated by the addition of an excess or overage of the active ingredient. Overages, then, are added to pharmaceutical formulations to keep the content of the active ingredient within the limits compatible with therapeutic requirements, for a predetermined period of time.

The amount of the overage depends upon the specific ingredient and the galenical dosage form. The International Pharmaceutical Federation has recommended that overages be limited to a maximum of 30% over the labeled potency of an ingredient.

PHARMACEUTICAL CONTAINERS

The official standards for containers apply to articles packaged by either the pharmaceutical manufacturer or the dispensing

pharmacist unless otherwise indicated in a compendial monograph. In general, repackaging of pharmaceuticals is inadvisable. However, if repackaging is necessary, the manufacturer of the product should be consulted for potential stability problems.

A pharmaceutical container has been defined as a device that holds the drug and is, or may be, in direct contact with the preparation. The immediate container is described as that which is in direct contact with the drug at all times. The liner and closure traditionally have been considered to be part of the container system. The container should not interact physically or chemically with the formulation so as to alter the strength, quality, or purity of its contents beyond permissible limits.

The choice of containers and closures can have a profound effect on the stability of many pharmaceuticals. Now that a large variety of glass, plastics, rubber closures, tubes, tube liners, etc are available, the possibilities for interaction between the packaging components and the formulation ingredients are immense. Some of the packaging elements themselves are subject to physical and chemical changes that may be time-temperature dependent.

Frequently, it is necessary to use a well-closed or a tight container to protect a pharmaceutical product. A *well-closed container* is used to protect the contents from extraneous solids or a loss in potency of the active ingredient under normal commercial conditions. A *tight container* protects the contents from contamination by extraneous materials, loss of contents, efflorescence, deliquescence, or evaporation and is capable of tight re-closure. When the packaging and storage of an official article in a well-closed or tight container is specified, water-permeation tests should be performed on the selected container.

In a stability program, the appearance of the container, with special emphasis on the inner walls, the migration of ingredients onto/into the plastic or into the rubber closure, the migration of plasticizer or components from the rubber closure into the formulation, the possibility of two-way moisture penetration through the container walls, the integrity of the tac-seal, and the back-off torque of the cap, must be studied.

GLASS—Traditionally, glass has been the most widely used container for pharmaceutical products to ensure inertness, visibility, strength, rigidity, moisture protection, ease of reclosure, and economy of packaging. While glass has some disadvantages, such as the leaching of alkali and insoluble flakes into the formulation, these can be offset by the choice of an appropriate glass. As the composition of glass may be varied by the amounts and types of sand and silica added and the heat treatment conditions used, the proper container for any formulation can be selected.

According to USP 26, glass containers suitable for packaging pharmacopeial preparations may be classified as either Type I, Type II, Type III, or type NP. Containers of Type I borosilicate glass are generally used for preparations that are intended for parenteral administration, although Type II treated soda-lime glass may be used where stability data demonstrates its suitability. Containers of Type III and Type NP are intended for packaging articles intended for oral or topical use.

New, unused glass containers are tested for resistance to attack by high-purity water by use of a sulfuric acid titration to determine the amount of released alkali. Both glass and plastic containers are used to protect light-sensitive formulations from degradation. The amount of transmitted light is measured using a spectrometer of suitable sensitivity and accuracy.

Glass is generally available in flint, amber, blue, emerald green, and certain light-resistant green and opal colors. The blue-, green-, and flint-colored glasses, which transmit ultraviolet and violet light rays, do not meet the official specifications for light-resistant containers.

Colored glass usually is not used for injectable preparations, since it is difficult to detect the presence of discoloration and particulate matter in the formulations. Light-sensitive drugs for parenteral use usually are sealed in flint ampuls and placed in a box. Multiple-dose vials should be stored in a dark place.

Manufacturers of prescription drug products should include sufficient information on their product labels to inform the pharmacist of the type of dispensing container needed to maintain the identity, strength, quality, and purity of the product. This brief description of the proper container, e.g., light- resistant, well-closed, or tight, may be omitted for those products dispensed in the manufacturer's original container.

PLASTICS—Plastic containers have become very popular for storing pharmaceutical products. Polyethylene, polystyrene, polyvinyl chloride, and polypropylene are used to prepare plastic containers of various densities to fit specific formulation needs.

Factors such as plastic composition, processing and cleaning procedures, contacting media, inks, adhesives, absorption, adsorption, and permeability of preservatives also affect the suitability of a plastic for pharmaceutical use. Hence, biological test procedures are used to determine the suitability of a plastic for packaging products intended for parenteral use and for polymers intended for use in implants and medical devices. Systemic injection and intracutaneous and implantation tests are employed. In addition, tests for nonvolatile residue, residue on ignition, heavy metals, and buffering capacity were designed to determine the physical and chemical properties of plastics and their extracts.

The high-density polyethylene (HDPE) containers, which are used for packaging capsules and tablets, possess characteristic thermal properties, a distinctive infrared absorption spectrum, and a density between 0.941 and 0.965 g/cm^3. In addition, these containers are tested for light transmission, water-vapor permeation, extractable substances, nonvolatile residue, and heavy metals. When a stability study has been performed to establish the expiration date for a dosage form in an acceptable high-density polyethylene container, any other high-density polyethylene container may be substituted provided that it, too, meets compendial standards and that the stability program is expanded to include the alternative container.

Materials from the plastic itself can leach into the formulation, and materials from the latter can be absorbed onto, into, or through the container wall. The barrels of some plastic syringes bind various pharmaceutical preservatives. However, changing the composition of the syringe barrel from nylon to polyethylene or polystyrene has eliminated the binding in some cases.

A major disadvantage of plastic containers is the two-way permeation or *breathing* through the container walls. Volatile oils and flavoring and perfume agents are permeable through plastics to varying degrees. Components of emulsions and creams have been reported to migrate through the walls of some plastics, causing either a deleterious change in the formulation or collapse of the container. Loss of moisture from a formulation is common. Gases, such as oxygen or carbon dioxide in the air, have been known to migrate through container walls and affect a preparation.

Solid dosage forms, such as penicillin tablets, when stored in some plastics, are affected deleteriously by moisture penetration from the atmosphere into the container.

Single unit does packaging in the form of blister packages are often used to package capsule and tablet dosage forms. A typical blister package is comprised of a polymeric film that is molded to have a cavity into which the dosage form is placed. The polymer film is then heat bonded to a paper backed foil liner.

As with plastic bottles, the blister package will allow a certain amount of moisture vapor permeation to occur, and this must be a consideration when selecting the type of film used for the package. The choice of packaging materials used depends on the degree to which the product needs to be protected from light, heat and moisture. Each material has different resistance to each of these elements and will affect the shelf life and storage conditions of the packaged pharmaceutical.

Polyvinylchloride (PVC) offers the least resistance to moisture vapor permeation. Polyvinylidenechloride (PVdC) has characteristics similar to PVC but offers superior resistance to moisture vapor permeation. Aclar, which is a polychlorotrifluo-

roethylene (PVC-CTFE) film has the lowest water vapor permeability and thus offers the best protection from moisture.

METALS—The pharmaceutical industry was, and to a degree still is, a tin stronghold. However, as the price of tin constantly varies, more-collapsible aluminum tubes are being used. Lead tubes tend to have pinholes and are little used in the industry.

A variety of internal linings and closure fold seals are available for both tin and aluminum tubes. Tin tubes can be coated with wax or with vinyl linings. Aluminum tubes are available with epoxy or phenolic resin, wax, vinyl, or a combination of epoxy or phenolic resin with wax. As aluminum is able to withstand the high temperatures required to cure epoxy and phenolic resins adequately, tubes made from this metal presently offer the widest range of lining possibilities.

Closure fold seals may consist of unmodified vinyl resin or plasticized cellulose and resin, with or without added color.

Collapsible tubes are available in many combinations of diameters, lengths, openings, and caps. Custom-use tips for ophthalmic, nasal, mastitis, and rectal applications also are available. Only a limited number of internal liners and closure seals are available for tubes fitted with these special-use tips.

Lined tubes from different manufacturers are not necessarily interchangeable. While some converted resin liners may be composed of the same base resin, the actual liner may have been modified to achieve better adhesion, flow properties, drying qualities, or flexibility. These modifications may have been necessitated by the method of applying the liner, the curing procedure, or, finally, the nature of the liner itself.

CLOSURES

The closures for the formulations also must be studied as a portion of the overall stability program. While the closure must form an effective seal for the container, the closure must not react chemically or physically with the product. It must not absorb materials from the formulation or leach its ingredients into the contents.

The integrity of the seal between the closure and container depends on the geometry of the two, the materials used in their construction, the composition of the cap liner, and the tightness with which the cap has been applied. Torque is a measure of the circular force, measured in inch-pounds, which must be applied to open or close a container. When pharmaceutical products are set up on a stability study, the formulation must be in the proposed market package. Thus, they should be capped with essentially the same torque to be used in the manufacturing step.

Rubber is a common component of stoppers, cap liners, and parts of dropper assemblies. Sorption of the active ingredient, preservative, or other formulation ingredients into the rubber and the extraction of one or more components of the rubber into the formulation are common problems.

The application of an epoxy lining to the rubber closure reduces the amount of leached extractives but essentially has no effect on the sorption of the preservative from the solution. Teflon-coated rubber stoppers may prevent most of the sorption and leaching.

REFERENCES

1. Lintner CJ. *Quality Control in the Pharmaceutical Industry,* vol 2. New York: Academic, 1973, p 141.
2. Connors KA, Amidon GL, Stella JV. *Chemical Stability of Pharmaceuticals,* 2nd ed. New York: Wiley, 1986.
3. Carstensen, JT. *Drug Stability Principles and Practices*. New York: Marcel Dekker, 1990
4. *Current Good Manufacturing Practice, 21 CFR 211.*
5. USP 26, 2003
6. *Guideline for Submitting Documentation for the Stability of Human Drugs and Biologics*. FDA, Center for Drugs and Biologics. Office of Drug Research Review, Feb 1987.
7. Lintner CJ, et al. *Am Perfum Cosmet* 1970; 85(12):31.
8. Kennon L. *J Pharm Sci* 1964; 53:815.
9. McMinn CS, Lintner CJ. (Oral presentation), APhA Acad Pharm Sci Mtg Ind Pharm Tech Sec. Chicago, May 1973.
10. Bentley DL. *J Pharm Sci* 1970; 59:464.

BIBLIOGRAPHY

Analysis. San Diego, Academic Press, 2001, Chap 13.

Carstensen, JT. *Drug Stability: Principles and Practices*, 2nd ed. New York: Marcel Dekker, 1995.

Cha J, Ranweiler JS, Lane PA. Stability studies. In Ahuja S, Scypinski S, eds. *Handbook of Modern Pharmaceutical Analysis.* San Diego: Academic Press, 2001.

Connors KA, Amidon GL, Stella VJ. *Chemical Stability of Pharmaceuticals*. New York: Wiley, 1986.

Documentation Practices: A Complete Guide to Document Development and Management for GMP and ISO9000 Compliant Industries. C DeSain, Advanstar Comm Inc, 1998.

Florence AT, Attwood D. *Physicochemical Principles of Pharmacy*, 2nd ed. New York: Chapman and Hall, 1988, Chap 4.

Florey K. *STP Pharma* 1986; 2:236.

Grimm W, Krummen K. *Stability Testing in the EC, Japan and the USA*. Stuttgart: Wiss. Verl.-Ges, 1993.

ICH Q1A (R): Stability Testing of New Drug Substances and Products. Step 4 Draft, 2003.

ICH Q1B: Photostability Testing of New Drug Substances and Products, 1996.

ICH Q1C: Stability Testing of New Dosage Forms, 1996.

ICH Q1D: Bracketing and Matrixing Designs for Stability Testing of New Drug Substances and Products. Step 4, 2003.

ICH Q1E: Evaluation OF Stability Data. Step 4, Draft, 2003.

ICH Q1F: Stability Data Package for Registration Applications in Climatic Zones III and IV, 2003.

Irwin WJ. *Kinetics of Drug Decomposition: Basic Computer Solutions*. Amsterdam: Elsevier, 1990.

Lachman L, et al. *The Theory and Practice of Industrial Pharmacy,* 3rd ed. Philadelphia: Lea & Febiger, 1986.

USP 24, Section <1077>, 1999.

Wagner JG, ed. *Biopharmaceutics and Relevant Pharmacokinetics.* Hamilton, IL: Hamilton Press, 1971.

Wells, JI. *Pharmaceutical Preformulation: The Physicochemical Properties of Drug Substances*. Chinchester: Ellis Horwood, 1988, Chap 5.

Windheuser JJ, ed. *The Dating of Pharmaceuticals*. Madison, WI: University Extension, University of Wisconsin, 1970.

CHAPTER 53

Bioavailability and Bioequivalency Testing

Henry J Malinowski, PhD
Steven B Johnson, PharmD

Oral solid dosage forms, tablets and capsules, are prescribed widely and are a very effective means of providing drugs to patients. A basic assumption is that when an oral solid dosage form is used by a patient, the drug from the dosage form is released, dissolves, and is absorbed promptly and consistently. Drug product quality is needed for this to be a valid assumption. In addition, many drugs are incompletely absorbed, due to factors relating to the drug, dosage form, and human physiology in the gastrointestinal tract. Optimal and consistent absorption of such drugs needs to be assured. Bioavailability and bioequivalence become important considerations in assuring optimal drug absorption. Major aspects of these areas are the topics covered in this chapter.

The bioavailability of a drug in an oral solid dosage form can be affected by numerous factors including food, changes in the metabolism of the drug due to drug interactions, gastrointestinal transit time, and changes in release characteristics of the drug from the dosage form (especially for modified release products). Changes in bioavailability can be thought of in terms of changes in exposure to the drug, which, if substantial, can relate to safety and efficacy concerns.

Bioequivalence is an important consideration in several key situations involving lot to lot consistency, innovator to generic product therapeutic equivalence, and situations where a marketed product undergoes changes in certain aspects including formulation, manufacturing process, and dosage strength.

In this chapter, bioavailability and bioequivalence topics are emphasized. Chemical equivalence, lot-to-lot uniformity of physicochemical characteristics, and stability equivalence are other factors that are important, as they too can affect product quality.

GENERAL CONCEPTS

Regarding bioavailability and bioequivalence, it is best to start with the basic concepts and factors that can affect the bioavailability of a drug and consider how these can influence bioequivalence and the clinical outcome of drug treatment. At the outset, the terms used in this chapter require careful definition, since, as in any area, some terms have been used in many different contexts by different authors.

Bioavailability is a term that indicates measurement of both the rate of drug absorption and total amount (extent) of drug that reaches the general circulation from an administered dosage form. It is specific to the active drug substance as contrasted to metabolites.

Equivalence is more a general, relative term that indicates a comparison of one drug product with another or with a set of established standards. Equivalence may be defined in several ways:

Chemical equivalence indicates that two or more dosage forms contain the same labeled quantities (plus or minus specified range limits) of the drug.

Clinical equivalence occurs when the same drug from two or more dosage forms gives identical *in vivo* effects as measured by a pharmacological response or by control of a symptom or disease.

Therapeutic equivalence implies that two brands of a drug product are expected to yield the same clinical result. The FDA specifically uses the term therapeutic equivalence in the evaluation of multisource prescription drug products.

Bioequivalence indicates that a drug in two or more similar dosage forms reaches the general circulation at the same relative rate and the same relative extent (ie, that the plasma level profiles of the drug obtained using the two dosage forms are the same).

Pharmaceutical equivalence refers to two drug products with the same dosage form and same strength.

THERAPEUTIC EQUIVALENCE EVALUATIONS

The FDA publication *Approved Drug Products with Therapeutic Equivalence Evaluations* identifies drug products approved on the basis of safety and effectiveness. In addition, this list contains therapeutic equivalence evaluations for approved multisource prescription drug products. These evaluations have been prepared to serve as public information and advice to state health agencies, physicians, and pharmacists to promote public education in the area of drug product selection and to foster containment of health-care costs.

To help contain drug costs, virtually every state has adopted laws and/or regulations that encourage the substitution of drug products. These state laws generally require either that substitution be limited to drugs on a specific list (the positive formulary approach) or that substitution be permitted for all drugs except those prohibited by a particular list (the negative formulary approach). Because of the number of requests for FDA assistance in preparing both positive and negative formularies, it became apparent that the FDA could not serve the needs of each state on an individual basis. The agency also recognized that providing a single list based on common criteria would be preferable to evaluating drug products on the basis of differing definitions and criteria in various state laws. The therapeutic equivalence evaluations in this publication reflect FDA's application of specific criteria to the approved multisource prescription drug products.

FDA classifies as therapeutically equivalent those products that meet the following general criteria:

1. They are approved as safe and effective.
2. They are pharmaceutical equivalents in that they (1) contain identical amounts of the same active drug ingredient in the same dosage form and route of administration and (2) meet compendial

or other applicable standards of strength, quality, purity, and identity.
3. They are bioequivalent in that (1) they do not present a known or potential bioequivalence problem, and they meet an acceptable *in vitro* standard, or (2) if they do present such a known or potential problem, they are shown to meet an appropriate bioequivalence standard.
4. They are adequately labeled.
5. They are manufactured in compliance with Current Good Manufacturing Practice regulations.

This concept of therapeutic equivalency applies only to drug products containing the same active ingredient(s) and does not encompass a comparison of different therapeutic agents used for the same condition. The FDA considers drug products to be therapeutically equivalent if they meet the criteria outlined above, even though they may differ in certain other characteristics such as shape, scoring configuration, release mechanisms, packaging, excipients, expiration date/time, and minor aspects of labeling (eg, the presence of specific pharmacokinetic information). The FDA believes that products classified as therapeutically equivalent can be substituted with the full expectation that the substituted product will produce the same clinical effect and safety profile as the prescribed product.

Methods for Determining Bioequivalence

Bioequivalence usually involves human testing but sometimes may be demonstrated using an *in vitro* bioequivalence standard, especially when such an *in vitro* test has been correlated with human *in vivo* bioavailability data. In other situations, bioequivalence may sometimes be demonstrated through comparative clinical trials or pharmacodynamic studies.

The FDA has categorized (21CFR320.24) various *in vivo* and *in vitro* approaches that may be utilized to establish bioequivalence. In descending order of accuracy, sensitivity and reproducibility these are:

1. An *in vivo* test in humans in which the active drug substance, as well as active metabolites when appropriate, is measured in plasma.
2. An *in vitro* test that has been correlated with human *in vivo* bioavailability data. This approach is most likely for oral extended release products and is described in detail in an FDA Guidance.
3. An *in vivo* test in animals that has been correlated with human bioavailability data.
4. An *in vivo* test in humans, where urinary excretion of the active drug substance, as well as active metabolites when appropriate, is measured.
5. An *in vivo* test in humans in which an appropriate acute pharmacological effect is measured.
6. Well-controlled clinical trials in humans that establish the safety and efficacy of the drug product, for establishing bioavailability. For bioequivalence, comparative clinical trials may be considered. This approach is the least accurate, sensitive, and reproducible approach and should be considered only if other approaches are not feasible.
7. A currently available *in vitro* test, acceptable to FDA, that ensures bioavailability. This approach is intended only when *in vitro* testing is deemed adequate, but no *in vitro in vivo* correlation has been established. It also can relate to considerations involving the Biopharmaceutics Classification System (BCS).

Most bioequivalence studies involve the direct measurement of the parent drug, as described in (1) above. Bioequivalence testing in animals is not a recommended approach due to possible differences in metabolism, gastrointestinal physiology, weight, and diet.

Minimizing the Need for Bioequivalence Studies

To minimize the need for human testing for bioequivalence, various approaches have been suggested:

1. *Situation where no changes are made for an approved, marketed product*–If a drug product has been adequately tested and approved for marketing, and if no changes in the manufacturing of the product are made, it is reasonable to assume that all subsequent batches of the product would be expected to be bioequivalent to the original product. If subsequently manufactured batches meet all tests of quality, including the dissolution test, no further human bioequivalence testing is needed.
2. *Situation where changes are made for an approved, marketed drug product*–Depending on the degree of change, bioequivalence may sometimes need to be reconfirmed. Although it is somewhat difficult to categorize such major changes, this issue has been addressed in a series of FDA Guidances Related to Scale-Up and Post-Approval Changes (SUPAC).
3. *Situation where human bioequivalence testing may not be needed for initial approval or for major post-approval changes*–Drug characteristics related to solubility and permeability may allow a reasonable expectation that the drug is unlikely to be subject to significant bioavailability problems. For such drugs, *in vitro* dissolution testing may be adequate, in lieu of *in vivo* testing. These concepts are described in the Biopharmaceutics Classification System (BCS). This classification system provides a scientific framework for classifying drugs based on aqueous solubility and intestinal permeability. In addition, criteria for rapid dissolution are described (not less than 85% dissolved in 30 minutes, using mild agitation and physiological media). The BCS permits waivers of in vivo bioequivalence testing for high solubility, high permeability drugs (Class I), which are formulated into immediate release dosage forms having rapid dissolution. A basic concept behind the BCS is that solutions of drugs are thought to have few bioavailability or bioequivalence issues. Dosage forms that contain high solubility drugs that exhibit rapid dissolution behave similar to solutions when either a solution or the highly soluble drug is in the stomach. Particularly for such drugs that are, in addition, highly permeable (well absorbed), the likelihood of bioavailability issues is quite small and, consequently, in vivo bioequivalence testing for such drugs is thought to be unnecessary. Similarly for oral solutions, in vivo bioequivalence testing is not necessary.

Additional Information

A significant factor related to drug bioavailability is the fact that many times drug is administered not as a solution but as a solid dosage form. Optimal bioavailability might be expected from a solution, since drug must first dissolve to be absorbed, but considerations such as drug stability, unpalatable taste, and desired duration of action (for controlled-release drug products) may prevent the use of a solution dosage form.

DOSAGE FORMS—In the dose titration of any patient the objective is, in conceptual terms, to attain and maintain a blood level that exceeds the minimum effective level required for response but does not exceed the minimum toxic (side-effect) level. This is shown graphically in Figure 53-1. There are several major absorption factors that can affect the general shape of this blood-level curve and thus drug response.

The Dose of the Drug Administered—The blood levels will rise and fall in proportion to the dose administered.

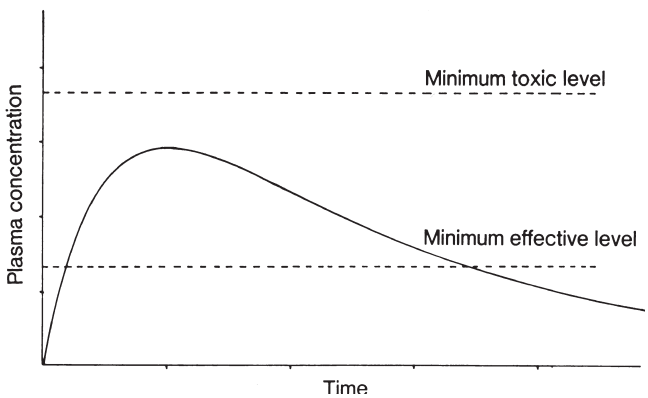


Figure 53-1. Typical plasma-level curve of a drug with effective and toxic (side-effect) profile levels defined.

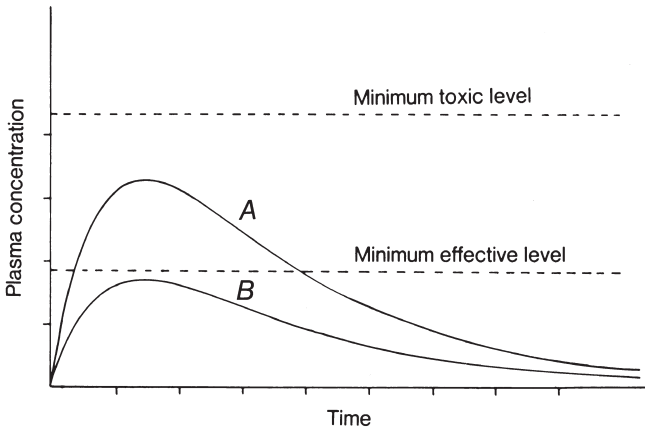


Figure 53-2. Effect of the extent of drug absorption from a dosage form on drug-plasma levels and efficacy. The extent of absorption from dosage form *B* is 50% of that from dosage form *A*.

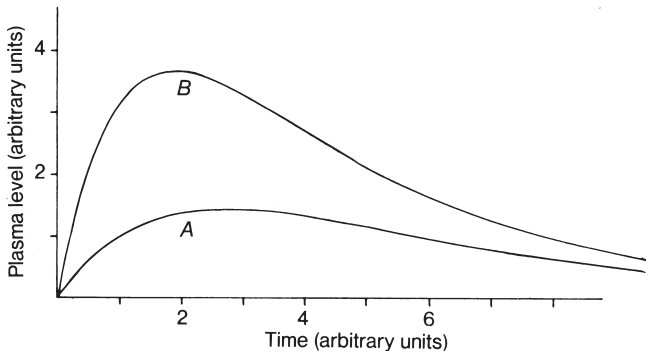


Figure 53-4. Computer simulation of the plasma-level curves for two dosage forms of the same drug assuming that the rate and extent of drug absorption for dosage form *A* were 50% and 50%, respectively, of those for dosage form *B*.

The Amount of Drug Absorbed from a Given Dosage Form—This involves the same principle as the first factor but is brought about by a different process. The effect of having only one half of the drug absorbed from a dosage form is equivalent to lowering the dose (Fig 53-2).

The Rate of Absorption of the Drug—If absorption from the dosage form is more rapid than the rate of absorption that gave the profile in Figure 53-1, minimum toxic (side-effect) levels may be exceeded. If absorption from the dosage form is sufficiently slow, minimum effective levels may never be attained (Fig 53-3).

A Combination of These Last Two Factors—This is also possible (Fig 53-4) and is probably the most likely situation in real life.

In any of these instances, the time course and extent of clinical response to the drug may be altered because of changes in dose or rate and extent of absorption.

Both factors, rate and extent of drug absorption, can be affected by the dosage form in which the drug is contained. The effect of rate of absorption may be intentional, as in controlled-release products, or unintentional, as brought about by, for example, a change in the composition and/or method of manufacture of the dosage form.

The choice of the inactive ingredients (excipients) used to prepare a dosage form is up to the individual manufacturer. It is through these changes, in composition and manufacturing technique, that unintended changes in bioavailability and bioequivalence may occur. Revalidation of bioequivalence may be needed for major changes in the manufacturing process, whereas small changes may not raise significant bioavailability concerns. In situations involving minor changes in the manufacturing process, comparative dissolution testing of the original and reformulated product provides adequate documentation of continued product quality, if the resulting dissolution profiles are similar. These considerations apply to all drug manufacturers, both innovator and generic companies. A description of the formulation of dosage forms and the factors that must be considered is given in Chapter 38.

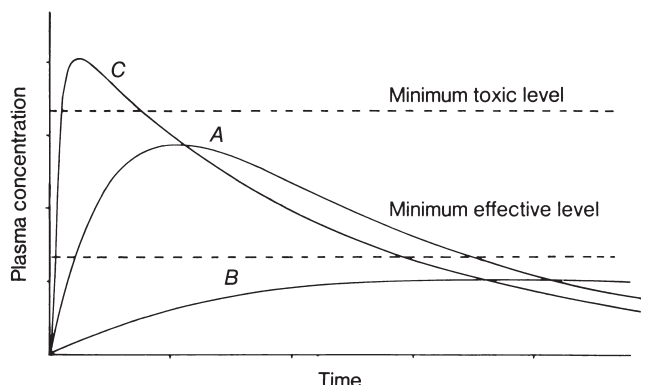


Figure 53-3. Effect of the rate of drug absorption from a dosage form on the plasma-level profile and efficacy. The rates of absorption from dosage forms *B* and *C* and 1/10 and 10 times those from dosage form *A*.

DISSOLUTION—For a drug to be absorbed, it must first go into solution. In Figure 53-5, the steps in the dissolution and absorption of a tablet or capsule dosage form are outlined. Similar profiles could be obtained for any solid or semisolid dosage form, including oral suspensions, parenteral suspensions, and suppositories. The theory and mechanics of drug dissolution rate are described in detail in Chapter 35. The physical characteristics of the drug and the composition of the tablet (dosage form) can have an effect on the rates of disintegration, deaggregation, and dissolution of the drug. As such, these can affect the rate of absorption and resultant blood levels of the drug.

An important aspect of product quality for marketed oral solid dosage forms relates to dissolution testing. Nearly all of the dosage forms actually used by patients will be from lots that have not directly undergone human bioavailability testing. It is previous batches of these products that would have been tested in humans. Once adequate product quality has been established by bioavailability testing, subsequent batches manufactured using the same formulation, equipment, and process are likely to be bioequivalent to the original batch tested in humans. This is an important concept in the regulatory control of product quality and is where *in vitro* testing such as assay, content uniformity, tablet hardness, and dissolution are involved. Among these several *in vitro* tests, dissolution testing is probably the most important, related to bioavailability. As part of the drug approval process, a dissolution test procedure is established for all oral solid dosage forms. These dissolution tests are incorporated in the

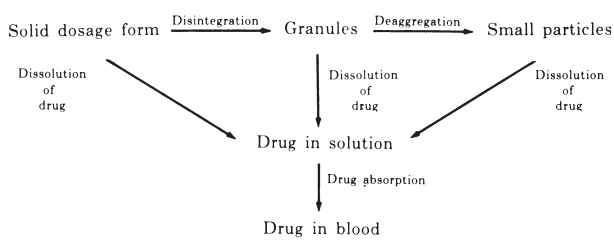


Figure 53-5. Sequence of events involved in the dissolution and absorption of a drug from a solid oral dosage form.

United States Pharmacopeia (USP) and apply both to innovator and generic drug products. All marketed batches of these drug products must meet the Abbreviated New Drug Application (ANDA)/New Drug Application (NDA)/USP dissolution tests throughout the shelf life of the product. Products failing their approved dissolution test and/or a USP dissolution test must be removed from the market.

Properties of the Drug—The physical characteristics of the drug that can alter bioavailability are discussed in Chapters 38 and 57 and consist of the polymorphic crystal form, choice of the salt form, particle size, use of the hydrated or anhydrous form, wettability, and solubility of the drug. Chapter 38 also discusses several other properties that can affect drug product quality adversely. Many of these factors should be discovered during the testing of the drug product prior to the marketing of the dosage form and should not, therefore, affect unknowingly the bioavailability of the drug product.

Properties of the Dosage Form—The various components of the solid or semisolid dosage form, other than the active ingredient, are discussed in Chapter 45. Only an overview, for tablet dosage forms, is given here. In addition to the active ingredient, a tablet product usually will contain the following types of inactive ingredients.

Glidants are used to provide a free-flowing powder from the mix of tablet ingredients, so that the material will flow when used on a tablet machine.

Binders provide a cohesiveness to the tablet. Too little binder will produce tablets that do not maintain their integrity; too much may affect adversely the release (dissolution rate) of the drug from the tablet.

Fillers are used to give the powder bulk so that an acceptably sized tablet is produced. Most commercial tablets weigh from 100 to 500 mg, so it is obvious that for many potent drugs the filler constitutes a large portion of the tablet. Binding of drug to the fillers may occur and affect bioavailability.

Disintegrants are used to cause the tablets to disintegrate when exposed to an aqueous environment. Too much will produce tablets that may disintegrate in the bottle because of atmospheric moisture; too little may be insufficient for disintegration to occur and may thus alter the rate and extent of release of the drug from the dosage form.

Lubricants are used to enhance the flow of the powder through the tablet machine and to prevent sticking of the tablet in the die of the tablet machine after the tablet is compressed. Lubricants are usually hydrophobic materials such as stearic acid, magnesium, or calcium stearate. Too little lubricant will not permit satisfactory tablets to be made; too much may produce a tablet with a water-impervious hydrophobic coat, which can inhibit the disintegration of the tablet and dissolution of the drug.

BIOEQUIVALENCE TESTING

The awareness of the potential for clinical differences between otherwise chemically equivalent drug products has been brought about by a multiplicity of factors that include, among others, better methods for clinical efficacy evaluation, development of techniques to measure microgram or nanogram quantities of drugs in biological fluids, improvements in the technology of dosage form formulation and physical testing, awareness of reported clinical inequivalencies in the literature, increased costs of classical clinical evaluation, the objective and quantitative nature of bioavailability tests, and the increase in the number of chemically equivalent products on the market because of patent expirations on the wonder drugs of the 1950s and 1960s as well as the Drug Price Competition and Patent Term Restoration Act of 1984, which established the generic drug approval procedures that are in place today.

The increase in the number of drugs that are available from multiple sources frequently has placed people involved in the delivery of health care in the position of having to select one from among several marketed products. As with any decision, the more pertinent the data available, the more comfortable one is in arriving at the final decision. The need to make these choices, in light of the potential for *in vivo* inequivalency among products or different batches of a given product, has increased the demand for quantitative data on the therapeutic equivalence of similar drug products. Bioequivalence testing represents one alternative solution to clinical testing for efficacy and is the means by which generic drugs are approved for marketing as well as the means by which the product quality of all drug products is maintained in situations involving major changes in formulation or manufacturing process.

Requirements for bioequivalence data on drug products should be applied reasonably. For example, with single-supplier drugs, bioequivalence testing is not an issue as far as brand-switching but can be a means of assessing changes between clinical and to-be-marketed formulations. In this context, the reason for bioequivalence testing should not be forgotten (ie, it is used as a surrogate, in certain situations, for the clinical evaluation of drug products). Bioequivalence data cannot be required if bioanalytical methodology is not available. However, in a number of cases, pharmacodynamic data may provide a more sensitive, objective evaluation of a product's therapeutic equivalence than clinical testing, and this can be an alternative approach in the absence of bioanalytical methodology.

Basic pharmacokinetic evaluation of bioavailability data is not necessary to show bioequivalence of two drug products. Pharmacokinetics has its major utility in the prediction or projection of dosage regimens and/or in providing a better understanding of observed drug reactions or interactions that result from the accumulation of drug in some specific site, tissue, or compartment of the body. The basis of the conclusion that two drug products are bioequivalent must be that the responses observed (blood, serum or plasma level, urinary excretion, or pharmacological response) for one drug product are essentially the same as the responses observed for the second drug product. The easy, but relatively rare, decisions in the evaluation of the bioequivalence of two drug products are those in which the two products are exactly superimposable (definitely bioequivalent) and those in which the two products differ in their bioequivalence parameters by a large amount, such as 50% or more (definitely not bioequivalent). Statistical evaluation of the data is necessary for all situations, particularly for data between these two extremes.

Evaluation of Bioequivalence Data

The following sections highlight some of the tests that should be considered when evaluating data from bioequivalence studies. The topics discussed are directed specifically toward plasma level evaluations. With minor modifications, the approaches outlined can be used for urinary excretion measurements or for suitable, quantitative, pharmacological response measurements.

Bioequivalence studies are usually conducted in healthy adults under standardized conditions. Most often, single doses of the test and reference product will be evaluated. However, in selected cases, multiple-dose regimens may be used (eg, when patients are used and they cannot be discontinued from a medication). The goal of the study is to evaluate the *in vivo* performance, as measured by rate and extent of absorption, of the dosage forms under standardized conditions, to minimize patient-related and other variability.

The protocol should define the acceptable age and weight range for the subjects to be included in the study as well as the clinical parameters that will be used to characterize a healthy adult (eg, physical examination observations, clinical chemistry, and hematological evaluations). The subjects should have been drug-free for at least 2 weeks prior to testing to eliminate possible drug-induced influences on liver enzyme systems. Normally, the subjects will fast overnight for at least 10 hours prior to dosing and will not eat until a standard meal is provided 4 hours post-dosing. The dosage forms should be given to subjects in a randomized manner, using a suitable crossover design, so that possible daily variations are distributed equally between the dosage forms tested. The protocol should define sample collection times and techniques to collect the biological fluid. The method of sample storage should also be defined.

Bioequivalence Assessment and Data Evaluation

Several parameters are used to provide a general evaluation of the overall rate and extent of absorption of a drug. An analysis of all characteristics is required before one can implicate any one factor or parameter as indicating bioequivalence or lack of bioequivalence. It is implicit that the analytical methodology used for analysis of drug in the samples is specific, sensitive, and precise.

The plasma concentration-time curve is the focal point of bioequivalence assessment and is obtained when serial blood samples, taken after drug administration, are analyzed for drug concentration. The concentrations are plotted on the ordinate (*y*-axis) and the times after drug administration that the samples were obtained, on the abscissa (*x*-axis).

A drug product is administered orally at time zero, and the plasma drug concentration at this time clearly should be zero. As the product passes through the gastrointestinal (GI) system (stomach, intestine), it must go through the sequence of events depicted in Figure 53-5. As the drug is absorbed, increasing concentrations of the drug are observed in successive samples until the maximum concentration is achieved. This point of maximum concentration (C_{max}) is called the peak of the concentration-time curve. If a simple model describes the pharmacokinetics of the drug tested, the peak concentration represents approximately the point in time when absorption and elimination of the drug have equalized.

The section of the curve to the left of the peak represents the absorption phase (or absorption and distribution), during which absorption predominates over elimination. The section of the curve to the right of the peak is called the elimination phase, during which elimination predominates over absorption. It should be understood that elimination begins as soon as the drug appears in the bloodstream and continues until all of the drug has been eliminated. Elimination is classically the log-linear portion of the curve. Absorption continues for some period of time into the elimination phase, for as long as there is drug (in gradually decreasing amounts) available for absorption in the GI tract.

One must recognize that elimination of the drug includes all processes of elimination of the drug, involving urinary excretion as well as metabolism by various tissues and organs. The *efficiency* of metabolism and urinary excretion will determine the shape of the elimination phase of the curve.

Bioequivalence studies normally are performed in healthy, adult volunteers under rigid conditions of fasting and activity because the objective is to obtain quantitative information on the influence of pharmaceutical formulation variables on the drug product's absorption. Drug blood-level profiles, therefore, allow quantification of the rate and extent of drug absorption and are critical in establishing the comparative efficiency of two drug products in delivering the drug to the systemic circulation.

Suggestions that bioequivalence studies should be performed in a disease-state population are not tenable if the object of the study is to assess drug formulations, unless safety considerations prohibit administration of the drug to healthy volunteers. If, on the other hand, the purpose is to determine the effect of disease on the efficiency of absorption of the drug product, then one must use the disease-state population. The reasoning is obvious. To ensure that any differences observed in the drug blood-level profiles are attributable to formulation factors, as much as possible, one must hold all other variables constant (ie, food, activity, and state of disease).

One need not be limited to drug blood-level profiles, but in a similar manner, may obtain cumulative urinary drug amount–time profiles. Drug concentration is determined in the urine at specified time intervals, and the amount excreted per interval is determined by multiplying the concentration by the volume of urine obtained in that interval. The amounts per interval then are combined, and ultimately the total amount excreted in the urine is obtained. This value is analogous to the area under the blood concentration-time curve. However, one limitation to this method is that rate cannot be readily determined. A typical cumulative urinary drug amount–time profile for several nitrofurantoin products is presented in Figure 53-6.

In assessing the bioequivalence of drug products, one must quantitate the rate and extent of absorption, which can be determined by evaluating parameters, derived from the blood-level concentration-time profile. Three parameters describing a blood level curve are considered important in evaluating the bioequivalence of two or more formulations of the same drug. These are the peak-height concentration, the time of the peak concentration, and the area under the blood (serum or plasma) concentration-time curve.

PEAK-HEIGHT CONCENTRATION—The peak of the blood level-time curve represents the highest drug concentration achieved after oral administration. It is reported as an amount per volume measurement (eg, micrograms/milliliter, units/milliliter, or grams/100 mL). The importance of this parameter is illustrated in Figure 53-7, where the blood concentration-time curves of two different formulations of a drug are represented. A line has been drawn across the curve at 4 μg/mL. Suppose that the drug is an analgesic and 4 μg/mL is the minimum effective concentration (MEC) of the drug in blood. If the blood concentration curves in Figure 53-7 represent the blood levels obtained after administration of equal doses of two formulations of the drug and it is known that analgesia would not be produced unless the MEC was achieved or exceeded, it becomes clear that formulation A would be expected to provide pain relief, while formulation B, even though it is well absorbed regarding extent of absorption, might be ineffective in producing analgesia.

On the other hand, if the two curves represent blood concentrations following equal doses of two different formulations of the same cardiac glycoside, and 4 μg/mL now represents the

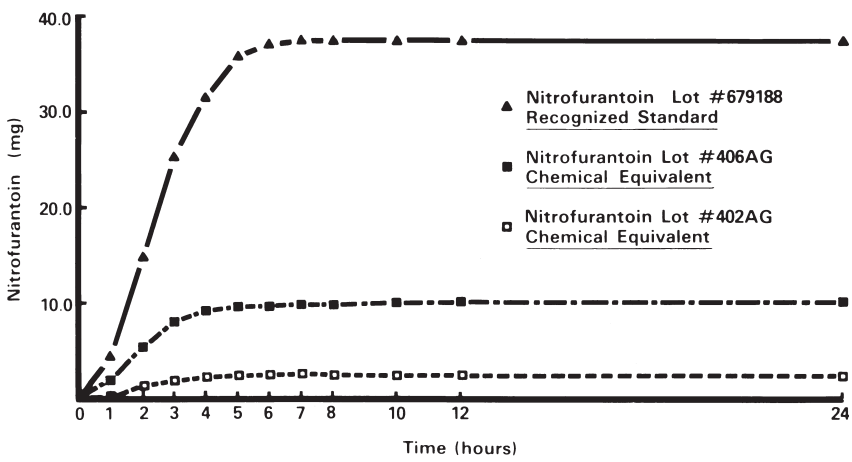


Figure 53-6. Average cumulative amounts of nitrofurantoin excreted form three lots of two commercially available products after a single oral dose of 100 mg of nitrofurantoin.

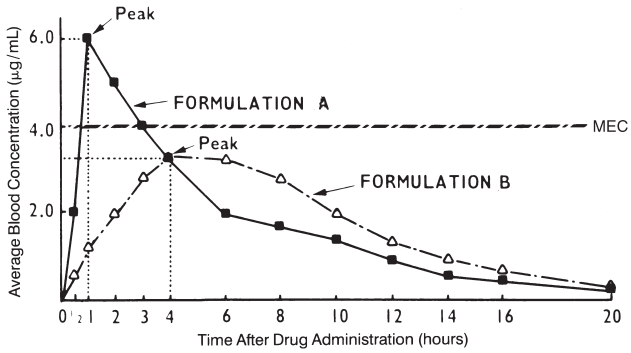


Figure 53-7. Blood concentration-time curves obtained for two different formulations of the same drug, demonstrating relationship of the profiles to the minimum effective concentration *(MEC)*.

minimum toxic concentration (MTC) and 2 μg/mL represents the MEC (Fig 53-8), formulation A, although effective, may also present safety concerns, while formulation B produces concentrations well above the MEC but never reaches toxic levels.

TIME OF PEAK CONCENTRATION—The second parameter of importance is the measurement of the length of time necessary to achieve the maximum concentration after drug administration. This parameter is called the time of peak blood concentration (T_{max}). In Figure 53-7, for formulation A, the time necessary to achieve peak blood concentration is 1 hour. For formulation B, T_{max} is 4 hours. This parameter is related closely to the rate of absorption of the drug from a formulation and may be used as a simple measure of rate of absorption but is normally not evaluated statistically.

To illustrate the importance of T_{max}, suppose that the two curves in Figure 53-8 now represent two formulations of an analgesic and that in this case the minimum effective concentration is 2 μg/mL. Formulation A will achieve the MEC in 30 minutes; formulation B does not achieve that concentration until 2 hours. Formulation A would produce analgesia much more rapidly than formulation B and would probably be preferable as an analgesic agent. On the other hand, if one were more interested in the duration of the analgesic effect than on the time of onset, formulation B would present more prolonged activity, maintaining serum concentrations above the MEC for a longer time (8 hours) than formulation A (5.5 hours).

AREA UNDER THE CONCENTRATION-TIME CURVE—The third, and sometimes the most important, parameter for evaluation, is the area under the serum, blood, or plasma concentration-time curve (AUC). This area is reported in amount/volume times time (eg, μg/mL × hr or g/100 mL × hr) and can be considered representative of the amount of drug absorbed following administration of a single dose of the drug.

Although several methods exist for calculating area under the concentration time curve (AUC), the trapezoidal rule method is most often used. This method assumes a linear function, $y = bt + a$, and its accuracy increases as the number of appropriate sampling intervals are increased. Table 53-1 and Figure 53-9 describes the process for calculating the AUC using the trapezoidal rule.

Returning to Figure 53-8, the curves, although much different in shape, have approximately the same areas (A = 34.4 μg/mL × hr; B = 34.2 μg/mL × hr), and both formulations can be considered to deliver the same amount of drug to the systemic circulation. Thus, one can see that AUC should not represent the only criterion on which bioequivalence is judged. All the results, as a composite, must be considered in reaching a decision about bioequivalence; no single parameter serves this purpose.

Table 53-1. Using the trapezoidal rule to calculate area under the concentration time curve. $AUC_{(0-\infty)}$ is used for bioequivalence analyses when the $AUC_{(0-t)}$ makes up ≥ 80% of the $AUC_{(0-\infty)}$. $AUC_{(0-t)}$ is used when the $AUC_{(0-t)}$ makes up < 80% of the $AUC_{(0-\infty)}$. When drugs with long half-lives, such as levothyroxine, are evaluated, $AUC_{(0-t)}$ is used and is truncated at 48 or 72 hours.

Area under the concentration-time curve from time zero to time t (AUC_{0-t})

1. Plot the concentration-time data for each subject;
2. Divide the curve into trapezoids by drawing vertical lines from each data point to the x-axis;
3. Calculate the area of the trapezoids using the following formula:
   - $AUC_{(t2-t1)} = [(C_2 + C_1)(t_2 - t_1)] / 2$
4. $AUC_{(0-t)}$ is then calculated by summing the individual areas to the time of the last concentration:
   - $AUC_{(0-t)} = AUC_{(t2-t1)} + AUC_{(t3-t2)} + \ldots + AUC_{(tn-(tn-1))}$

Area under the concentration-time curve from time zero to infinity ($AUC_{0-\infty}$)

5. To calculate $AUC_{(0-\infty)}$, the tail region of the curve must be added to the $AUC_{(0-t)}$:
   - $AUC_{(0-\infty)} = AUC_{(0-t)}$ + AUC "tail"
6. AUC "tail" = C_t / λ_z, where:
   - C_t is the last detectable concentration, and
   - λ_z is the terminal elimination rate constant (see Figure 58-9)

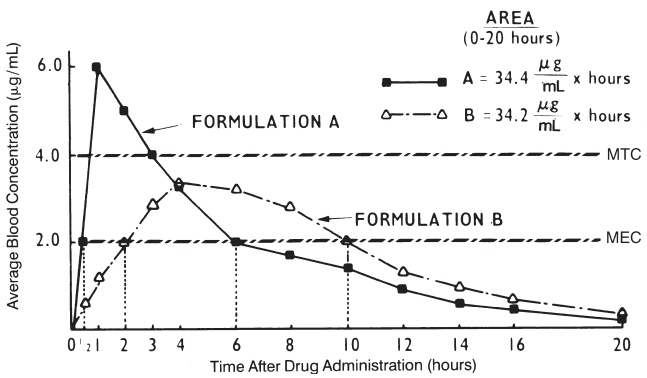


Figure 53-8. Blood concentration-time curves obtained for two different formulations of the same drug, demonstrating relationship of the profiles to the minimum toxic concentration *(MTC)* and the minimum effective concentration *(MEC)*.

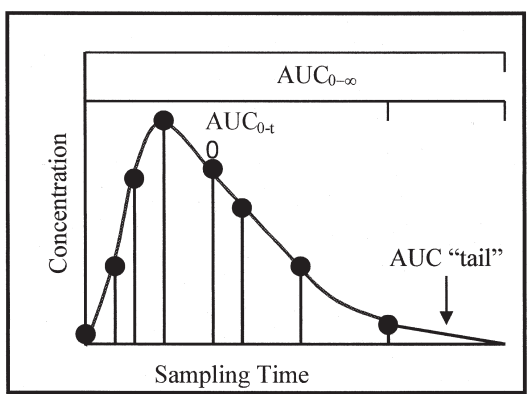


Figure 53-9. Graphical depiction using trapezoidal rule for calculating area under the concentration-time curve.

Criteria for Bioequivalence

Under the *Drug Price Competition and Patent Term Restoration Act of 1984,* manufacturers seeking approval to market a generic drug must submit data demonstrating that the drug product is bioequivalent to the pioneer (innovator) drug product. A major premise underlying the 1984 law is that bioequivalent drug products are therapeutically equivalent and, therefore, interchangeable.

The standard bioequivalence study is conducted in a crossover fashion in a small number of volunteers, usually with 24 to 36 healthy adults. The number of subjects appropriate for a bioequivalence study can be determined on the basis of previous knowledge of the drug's variability. In general, the number of subjects should be sufficient to detect 20% differences in the measured parameters, with 80% certainty. Single doses of the test and reference drugs are administered, and blood or plasma levels of the drug are measured over time. Characteristics of these concentration-time curves, such as the area under the curve (AUC) and the peak blood or plasma concentration (C_{max}) are examined by statistical procedures.

Bioequivalence of different formulations of the same drug substance involves equivalence with respect to the rate and extent of drug absorption. Two formulations whose rate and extent of absorption differ by ±20% or less are generally considered bioequivalent. The use of the ±20% criteria is based on a medical decision that for most drugs, a ±20% difference in the concentration of the active ingredient in blood will not be clinically significant.

To verify, for a particular pharmacokinetic parameter, that the ±20% criteria are satisfied, two one-sided statistical tests are carried out using the log-transformed data from the bioequivalence study. In order to interpret the statistical results, the log-transformed data must first be back-transformed. When the log-transformed data are back-transformed, the ±20% now becomes −20%/+25%. One test is used to verify that the lower bound of the 90% confidence interval of the average response for the generic product is no more than 20% below that of the innovator product; the other test is used to verify that the upper bound of the 90% confidence interval of the average response for the generic product is no more than 25% above that for the innovator product. The current practice is to carry out each of the two one-sided tests at the 0.05 level of significance.

Computationally, the two one-sided tests are carried out by computing a 90% confidence interval. For approval of ANDAs, in most cases, the generic manufacturer must show that a 90% confidence interval for the ratio of the mean response (usually AUC and C_{max}) of its product to that of the innovator is within the limits of 0.8 and 1.25, after the log-transformed data has been back-transformed. If the true average response of the generic product in the population is near 20% below, or 25% above, the innovator average, one or both of the confidence limits is likely to fall outside the acceptable range, and the product will fail the bioequivalence test. Thus, an approved product is likely to differ from the innovator by far less than this quantity. These same criteria are applied to other bioequivalence situations such as post-approval changes in innovator or generic products.

The current practice of carrying out two one-sided tests at the 0.05 level of significance ensures that if the two products truly differ by as much as or more than is allowed by the equivalence criteria, there is no more than a 5% chance that they will be approved as equivalent. This reflects the fact that the primary concern from the regulatory point of view is the protection of the patient against a conclusion of bioequivalence if this does not hold true. The results of a bioequivalence study usually must be acceptable for more than one pharmacokinetic parameter. As such, a generic product that truly differs by ±20% or more from the innovator product with respect to one or more pharmacokinetic parameters would have less than a 5% chance of being approved. Different statistical criteria may be used when bioequivalence is demonstrated through comparative clinical trials, pharmacodynamic studies, or comparative *in vitro* methodology.

Using the two one-sided test procedures, when two drug products differ by more than 12–13% in means, they are unlikely to pass the bioequivalence confidence interval criteria of 80–125%. A study of more than 200 approved generic drugs indicated a mean bioavailability difference of only 3.5% existed. Although somewhat larger differences might meet the bioequivalence criteria, the reality is that, for generic drug products approved by FDA, observed differences have been quite small.

Fed Bioequivalence Studies

Food has been shown to alter the bioavailability of some drugs, and this alteration can have a negative impact on the interpretation of bioequivalence results between test and reference products. As a result, bioequivalence studies are usually conducted under fasting conditions. However, in some instances a fasting study may not be reasonable for a particular drug because of safety considerations or perhaps because of the drug's intended clinical indication. In these situations, a fed bioequivalence study is sometimes acceptable. A fed bioequivalence study is similar to the standard bioequivalence study except that following an overnight fast, the test and reference products are administered 30 minutes after the start of a standardized meal. The FDA currently recommends a high-fat, high-calorie meal as described in an FDA Guidance. The composition of this meal is described in Table 53-2. The same statistical criteria, as used for the standard bioequivalence study, are observed for the resultant fed bioequivalence study data.

Average Bioequivalence

The standard *in vivo* bioequivalence study design is based on administration of the test and reference products on separate occasions to healthy subjects, either in single or multiple doses, with random assignment to the two possible sequences of drug product administration. Samples of plasma or blood are analyzed for drug and/or metabolite(s) concentrations, and pharmacokinetic parameters are obtained from the resulting concentration-time curves. Parameters are analyzed statistically to determine if the test and reference products yield comparable values. Statistical analysis for pharmacokinetic parameters, such as area under the curve (AUC) and peak concentration (C_{max}), is based on the two one-sided tests procedure, which determines whether the average values for pharmacokinetic parameters measured after administration

Table 53-2. The FDA Standardized High-Fat Test Meal Composition

The example test meal would be two eggs fried in butter, two strips of bacon, two slices of toast with butter, four ounces of hash brown potatoes, and eight ounces of whole milk. Substitutions in this test meal can be made as long as the meal provides a similar amount of calories from protein, carbohydrate, and fat and has a comparable meal volume and viscosity.

MEAL COMPOSITION	ENERGY (kcal)
Protein	150
Carbohydrate	250
Fat	500–600

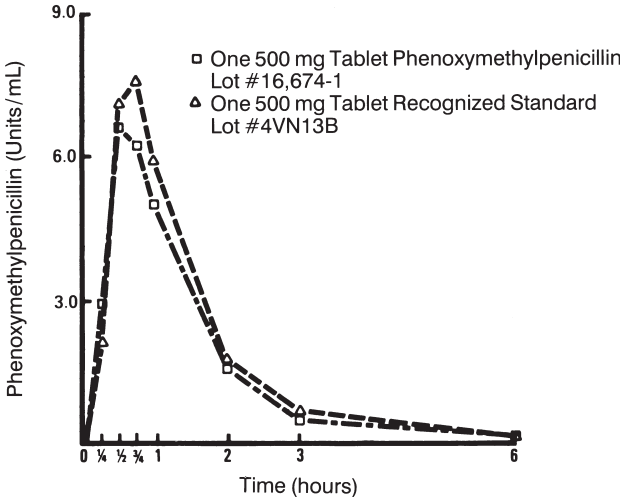


Figure 53-10. Average serum concentration of phenoxymethyl penicillin following oral administration of 500 mg given as one tablet of recognized standard (Δ) or of test product, research lot (□).

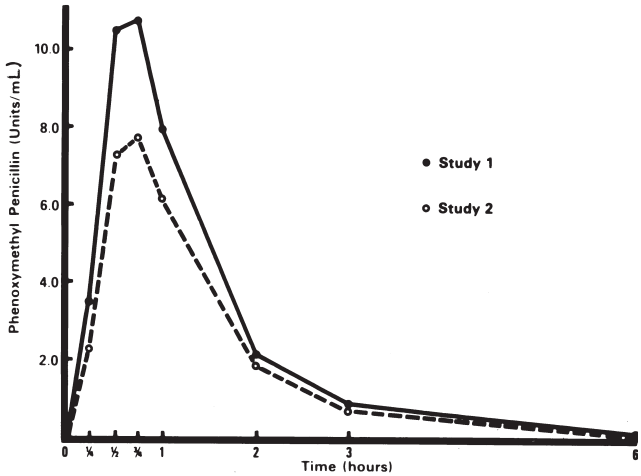


Figure 53-12. Average serum concentration of phenoxymethyl-pencillin following a single oral 500-mg dose of recognized standard, in two different subject populations.

of the test and reference products are comparable (ie, average bioequivalence). This procedure involves the calculation of a 90% confidence interval for the ratio of the averages of the test and reference product. To establish bioequivalence, the calculated confidence interval must fall within a bioequivalence limit, usually 80–125% for the ratio of the product averages. In addition to this general approach for determining bioequivalence, a 2001 FDA Guidance provides specific recommendations for (1) logarithmic transformation of pharmacokinetic data, (2) methods to evaluate sequence effects, and (3) methods to evaluate outlier data.

Population and Individual Bioequivalence

Statistically, the average bioequivalence approach focuses on the comparison of population averages of a bioavailability metric of interest and not on the variability of the metric for the test and reference products. In addition, average bioequivalence cannot describe a subject-by-formulation interaction, that is, the variation that may be present among individuals in the average test and reference difference. In contrast, population and individual bioequivalence approaches include comparisons of both averages and variability of the study metric. The population bioequivalence approach assesses the total variability of the metric in the population. The individual bioequivalence approach assesses the within-subject variability as well as the subject-by-formulation interaction. However, due to statistical and study design issues with population and individual bioequivalence, respectively, the FDA has deferred recommending these analyses methods.

Study Design

AVERAGE OR POPULATION BIOEQUIVALENCE—A conventional, nonreplicated crossover design, such as the standard two-formulation, two-period, two-sequence crossover design, may be used to generate data for assessment of population bioequivalence. Replicated-crossover designs or parallel designs also may be used.

INDIVIDUAL BIOEQUIVALENCE—Three important parameters, the within-subject variability for the test and reference metric and subject-by-formulation interaction variability components, are integral components of the individual bioequivalence criterion. A replicated-crossover design of the bioequivalence study should be used to estimate these parameters.

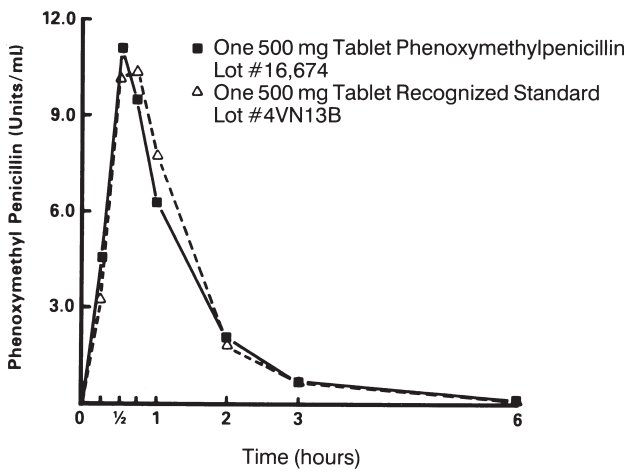


Figure 53-11. Average serum concentration of phenoxymethyl-pencillin following oral administration of 500 mg given as one tablet of recognized standard (Δ) or of test product full mfg lot (■).

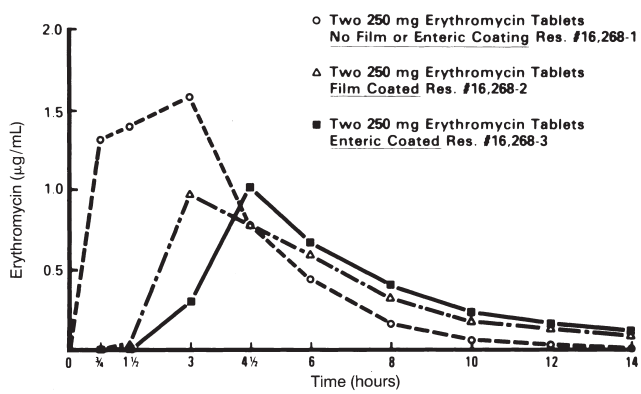


Figure 53-13. Average serum erythromycin concentration administered in 500-mg doses as three different tablet dosage forms. The results were obtained from 21 healthy adult subjects following an overnight fast of 12 hr before, and 2 hr after, drug administration.

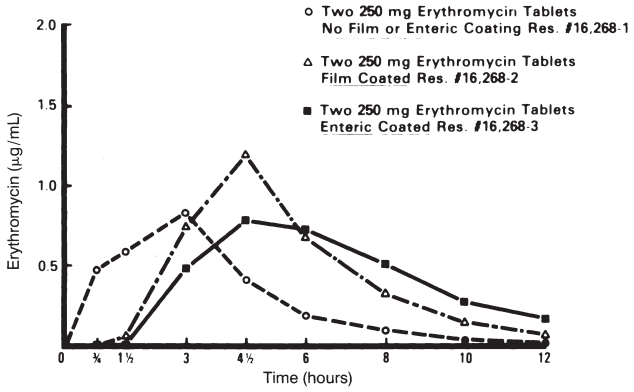


Figure 53-14. Average serum erythromycin concentration administered in 500-mg doses as three different tablet dosage forms. The results were obtained from 12 healthy adult subjects with only a 2-hr fast before drug administration.

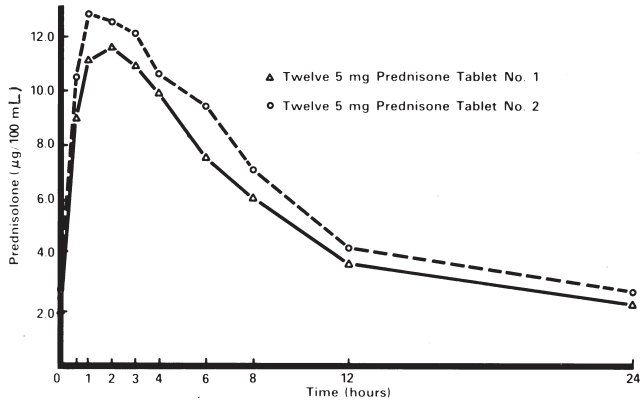


Figure 53-16. Average plasma prednisolone levels following 60 mg of prednisone administered to 24 normal adults as a single oral dose of 125-mg prednisone tablets from two different manufacturers. Plasma levels were determined by a competitive protein-binding assay.

Further information related to current FDA recommendations regarding the design and analysis of bioequivalence studies is available on the internet at http://www.fda.gov/cder/, under *Regulatory Guidance Documents*.

PITFALLS ASSOCIATED WITH CROSS-STUDY COMPARISONS—This is a situation in which the blood concentration-time curve of a drug product in one study is compared with the blood concentration-time curve of that drug product in another study. There are several reasons why such cross-study comparisons are not recommended and may lead to false conclusions. However, if no other data are available, and if important comparisons must be made, cross study comparison may be informative, keeping in mind the possible limitations. The following examples, used to illustrate these three points, are taken from actual bioavailability data.

Different Subject Population—In Figure 53-10, a research lot of potassium phenoxymethyl penicillin was compared with the appropriate reference standard for that product. The research lot drug was found to be bioequivalent, with average peak-serum concentrations differing by 8% and the area differing by only 9%. In another study conducted with a full-manufacture lot of the test product, the same lot of the reference standard potassium phenoxymethyl penicillin was used.

The results of this study are shown in Figure 53-11. Again, the two products were found to be bioequivalent, as the peak and area parameters differed by less than 5%. In these two studies, identical test conditions were used, and the same analytical procedure and laboratory was employed. However, if one compares the plasma levels for the reference standard lot found in Figure 53-10 with the levels for the same lot of tablets in the study in Figure 53-11, sizable differences in blood levels are found, as shown in Figure 53-12.

The average peak serum levels for this lot of tablets were found to be 8.5 and 12.5 units/mL in the two respective studies, a difference of approximately 31%. Likewise, the average AUC was found to differ by approximately 21%. Such apparent differences are solely the result of cross-study comparisons and are not due to differences in actual bioavailability.

The same lot of reference standard tablets was used in both studies. Hence, the difference must be due to the experimental variables that occur normally from study to study. The major difference between the two studies was the subject population involved. In the first study, healthy adult male prison volunteers were used, whereas in the second study, there were 17 females and 7 males in a hospital clinic, also described as normal, healthy volunteers. An appreciable difference in sex distribution was obvious when comparing these studies. Adjustments for body weight and surface area alone did not correct for the apparent discrepancies in peak concentration or blood level AUC. It is difficult to deter-

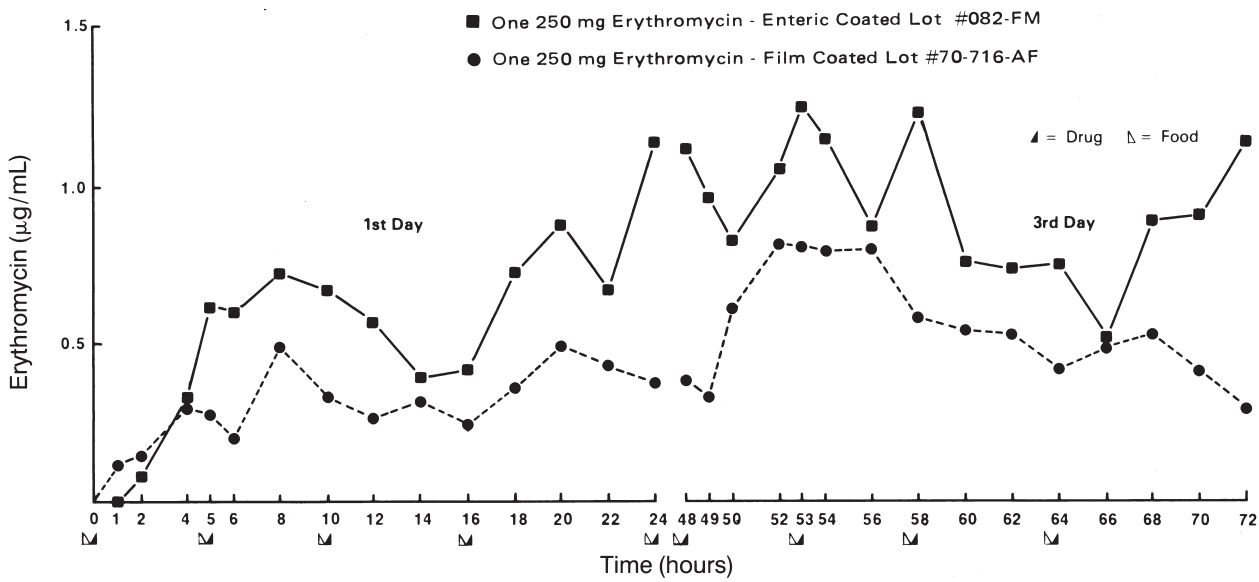


Figure 53-15. Average serum erythromycin concentration-time profiles from drug administered in two different tablet dosage forms. The results were obtained from 24 healthy adult subjects, following administration of 250 mg, four times a day, with meals and at bedtime.

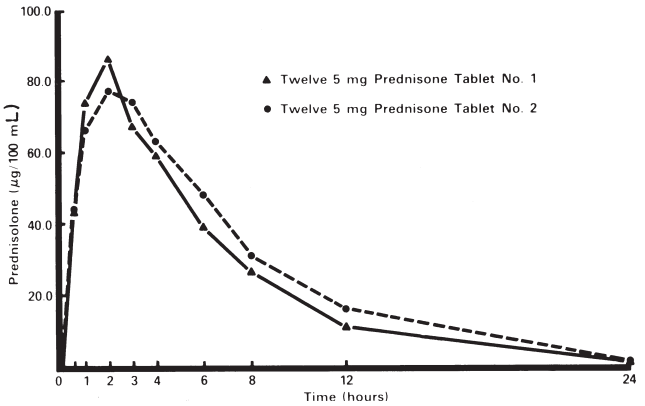


Figure 53-17. Average plasma prednisolone levels following 60 mg of prednisone administered to 24 normal adults as a single oral dose of 125-mg prednisone tablets from two different manufacturers. Plasma levels were determined by a radioimmunoassay procedure.

mine the exact factors that caused the observed differences. This example should serve as a note of caution in comparing bioavailability values of peak concentration and area under the curve from different studies.

Different Study Conditions—Parameters such as the food or fluid intake of the subject before, during, and after drug administration can have dramatic effects on the absorption of certain drugs. Figure 53-13 shows the results of a three-way crossover test in which the subjects were fasted 12 hours overnight and 2 hours after drug administration of an uncoated tablet, a film-coated tablet, or an enteric-coated tablet of erythromycin.

The results of this study suggest that the uncoated tablet is superior to both the film-coated and enteric-coated tablets in terms of blood level performance. These results also suggest that neither film coating nor enteric coating is necessary for optimal blood-level performance. Figure 53-14 shows results with the same tablets when the study conditions were changed to only a 2-hr preadministration fast with a 2-hr postadministration fast. In this case, the blood levels of the uncoated tablet were depressed markedly, while the film-coated and enteric-coated tablets showed relatively little difference in blood levels.

From this second study, it might be concluded that film coating appears to impart the same degree of acid stability as an enteric coating. This might be acceptable if only one dose of the antibiotics was required. However, Figure 53-15 shows the results of a multiple-dose study in which the enteric-coated tablet and the film-coated tablet were administered four times a day, immediately after meals. The results show that the film coating does not impart the degree of acid stability that the enteric coating does when the tablets are administered immediately after food in a typical clinical situation.

Different Assay Methodology—Depending on the drug under study, there may be more than one assay method available. For example, some steroids can be assayed by a radioimmunoassay, competitive protein-binding, gas-liquid chromatography, or indirectly by a 17-hydroxycorticosteroid assay.

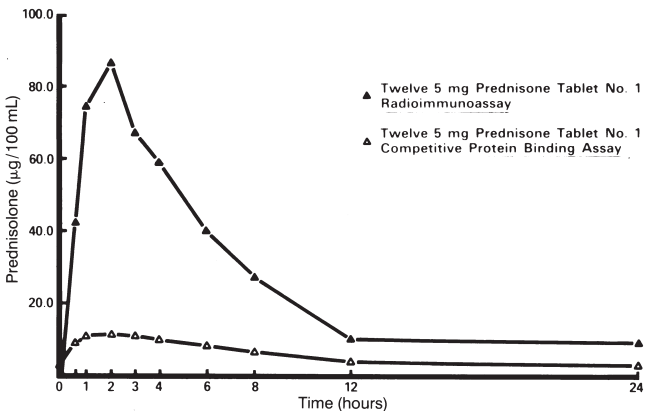


Figure 53-18. Average plasma prednisolone profiles from drug administered as a single 60-mg dose to 24 normal adults. Plasma levels were determined by both a competitive protein-binding assay and a radioimmunoassay.

Figures 53-16 and 53-17 show the results of a comparison of prednisone tablets using a competitive protein-binding method and a radioimmunoassay, respectively. The serum concentration-time curves resulting from each method lead to the same conclusion, that the products are bioequivalent. However, Figure 53-18 shows a comparison of the absolute values obtained by the two assay methods with the same product.

Obviously, the wrong conclusion would have been reached if one product had been assayed by one method and the other product by the other method and the results had been compared. Even in cases in which only one assay method is employed, there are numerous modifications with respect to technique among laboratories that could make direct comparisons difficult.

The backbone of any bioavailability study involving plasma (or urine) levels of drug, in addition to good study design and subject controls, is the analytical methodology used to determine the levels of a drug. In most cases, one probably can assume that the precision and reliability of the method employed in a given study have been established to a sufficient degree to make the results of the study internally consistent. As demonstrated, major problems arise when, without careful evaluation of the analytical methodology employed, one attempts to compare the data of studies from different laboratories. Even with similar analytical methodology performed by the same laboratory, it would be unreasonable to expect agreement, using the same dosage form, closer than 20% to 25% for plasma levels from one study to the next.

Under the best conditions, cross-study comparisons are relatively insensitive, and at worst they can be misleading. Cross-study comparisons certainly cannot be used to make decisions or estimate differences in drug products with the generally acceptable sensitivity of difference detection of 20% or less.

BIBLIOGRAPHY

Abdou HM. *Dissolution, Bioavailability and Bioequivalence*. Easton, PA: Mack Publishing Co, 1989.

Amidon GL, Robinson JR, Williams RL. *Scientific Foundations for Regulating Drug Product Quality*. Alexandria, VA: AAPS Press, 1997.

Chow S-C. *Design and Analysis of Bioavailability and Bioequivalence Studies*. New York: Dekker, 1992.

Marston SA, et al. Evaluation of direct curve comparison metrics applied to pharmaceutic profiles and relative bioavailability and bioequivalence. *Pharm Res* 14:1363, 1997.

Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the bioequivalence of average bioavailability. *J Pharmacokinet Biopharm* 1987; 15:657.

Shah VP, Maibach HI. *Topical Drug Bioavailability, Bioequivalence and Penetration*. New York: Plenum, 1993.

Shargel L, Yu AB. *Applied Biopharmaceutics and Pharmacokinetics*. Norwalk, CT: Appleton & Lange, 1993.

Welling PG, et al. *Pharmaceutical Bioequivalence.* New York: Dekker, 1991.

Guidance for Industry - Extended Release Oral Dosage Forms: *Development, Evaluation, and Application of In Vitro/In Vivo Correlations*. http://www.fda.gov/cder/guidance/1306fnl.pdf

Guidance for Industry–Immediate Release Solid Oral Dosage Forms–*Scale-Up and Postapproval Changes: Chemistry, Manufacturing, and Controls, In Vitro Dissolution Testing, and In Vivo Bioequivalence Documentation*. 1995. http://www.fda.gov/cder/guidance/cmc5.pdf

Guidance for Industry–*Questions and Answers about SUPAC-IR*. 1997. http://www.fda.gov/cder/guidance/qaletter.htm

Guidance for Industry–*SUPAC-IR/MR: Immediate Release and Modified Release Solid Oral Dosage Forms Manufacturing Equipment Addendum*. 1999. http://www.fda.gov/cder/guidance/1721fnl.pdf

Guidance for Industry–*SUPAC-MR: Modified Release Solid Oral Dosage Forms Scale-Up and Postapproval Changes: Chemistry, Manufacturing, and Controls; In Vitro Dissolution Testing and In Vivo Bioequivalence Documentation*. 1997. http://www.fda.gov/cder/guidance/1214fnl.pdf

Guidance for Industry–*Waiver of In Vivo Bioavailability and Bioequivalence Studies for Immediate-Release Solid Oral Dosage Forms Based on a Biopharmaceutics Classification System*. http://www.fda.gov/cder/guidance/3618fnl.pdf

Guidance for Industry–*Food-Effect Bioavailability and Fed Bioequivalence Studies*. 2002. http://www.fda.gov/cder/guidance/5194fnl.doc

NTI Letter. *Therapeutic Equivalence of Generic Drugs Response to National Association of Boards of Pharmacy*. 1997. http://www.fda.gov/cder/news/ntiletter.htm

CHAPTER 54

Plastic Packaging Materials

Barrett E Rabinow, PhD
Theodore J Roseman, PhD

As defined by the American Society for Testing and Materials (ASTM), a plastic is a material that contains as an essential ingredient one or more polymeric organic substances of large molecular weight, is solid in its finished state and at some stage in its manufacture or processing into finished articles can be shaped by flow. The large-molecular-weight organic substance is called a polymer.

The use of plastics in the health care industry has grown at a very rapid rate since the 1960s. This phenomenal growth is due primarily to the wide flexibility in choice of properties offered by plastics. However, because of the wide range of properties of plastics, judicious selection must be made for the intended application.

Prior to the recognition of the potential use of plastics in health care practice, glass was the predominate material used in the primary packaging of pharmaceutical products. Glass has a definite advantage in being a relatively unreactive and an inert substance (although leachable aluminum and glass particles or delamination have occasionally posed problems). As such, it can be used in contact with many critical products, either dry or liquid. It provides excellent protection against water vapor and gas permeation, and it can withstand steam sterilization (autoclaving) without incurring physical distortion. Two definite disadvantages of glass in the field of packaging, however, are its fragility and weight. Because of these negative aspects, coupled with the many positive attributes of plastics, significant inroads for the use of plastic in pharmaceutical packaging have been made. Today, for example, plastics are being used in the following primary packaging areas, where in the 1960s only glass could be considered: syringes, bottles, vials, and ampules.

There are many other significant medical uses that, without the use of plastics, would never have been technically feasible. A few examples include indwelling catheters, prosthetic devices, tracheotomy tubes, unit dose packaging, and flexible containers for intravenous, irrigation, and inhalation solutions, as well as for the collection of blood. An additional use for plastics is in secondary container packaging (ie, packaging that is not in direct contact with the product itself). This particular use normally involves plastic films of various types and thicknesses used for tamper-proof overwrapping, whereas the previously mentioned devices normally are fabricated by molding or extrusion of the finished part.

Selection of the appropriate materials for a packaging application should be performed with an understanding of the intended overall design of the package. The requirements should be specified with regard to customer usage, regulatory approval, marketing presentation, toxicological considerations, manufacturability, sterility, and, very importantly, protection of the pharmaceutical product or device during transportation, storage, and use. These functional requirements then must be analyzed in terms of the stress requirements they impose on the material, permitting translation of those requirements into material properties. A target material profile is developed by assigning required values of design and performance properties that predict or correlate with the container functions. Likely candidate materials are determined by comparing their properties with the property profile derived from the functional requirements. A prototype is built and tested via functionally oriented tests such as maintenance of product stability, simulated usage and storage tests, and customer focus groups. These concepts are embodied in ISO 11607.[1] Material properties affecting functional performance are described below.

MATERIAL PROPERTIES

Mechanical Properties

Important mechanical properties in plastic packaging materials are:

Tensile strength—the maximum force needed to pull apart a specimen of material, divided by its cross-sectional area. Elongation is the percentage change over original length at breaking point and measures a film's ability to stretch.

Impact strength—a measure of the ability to withstand shock-loading, in which a specimen receives a blow from a swinging pendulum, for example. Fracture will occur if the impact force exceeds the limit of elasticity of the material. Glass, for example, has a much lower impact strength than many plastics, although it has appreciable tensile strength.

Tear strength—measured both as the force necessary to initiate a tear and force to propagate a tear. Propagation of tear is undesirable in shipping sacks but desirable in tear tapes. Orientation of the material can affect results, because the polymer chains can be aligned along a particular direction during manufacturing, thus conferring greater strength in that direction.

Stiffness—the resistance of bending where deflection against a load can be measured.

Flex resistance to the development of pinholing and fracture, when subjected to repeated flexing or creasing, is important in shipping applications. Unsupported aluminum foil, unless it is heavy gauge, is prone to this failure mode.

Coefficient of friction or *slip*—relates to the ease with which one material will slide over another. Passage of films through

packaging machinery requires high slip to prevent binding and is important in form–fill-seal operations.

Blocking—the tendency of two adjacent layers of film to stick together. This can create difficulties during manufacturing.

Fatigue resistance, or the ability to withstand the imposition of repetitive short-time stress or deformation without cracking, is relevant in applications involving continual cyclical loading, such as toggle mechanisms, gear teeth of a pump, or peristaltic compression of IV tubing.

Creep failure occurs when a plastic is subjected to a constant static load; it deforms quickly and elastically (reversibly) to a predicted strain value and then continues to deform at a slower rate indefinitely. Rupture may eventually occur. Creep is both temperature and time dependent. The design life of the package thus plays a role, because both strength and stiffness may be time related. The loss of torque of a static bottle-closure system over time or deformation of plastic IV tubing under constant compression are examples.

Other properties of plastics may affect their usage in a particular application. For example, low-temperature mechanical behavior is important if a plastic is exposed to freezing temperatures during its use, since the impact strength of certain plastics decreases in the frozen state. The density of plastics, which varies between 0.8 and 1.8 g/cm^3, is an important property, since lower-density materials will produce more items per unit weight. Additionally, the melting point, which may extend over a range of temperatures, is important for determining processing temperatures, heat sterilizability, ability to hot-fire a product, and heat-sealing characteristics.

Additional mechanical properties are characteristic of component subsystems or of the entire package:

Hot tack—the ability of a heat seal to remain intact as it cools down from its sealing temperature, thus preserving package integrity.

Abrasion and shock test measures the interactive effects of abrasion and shock on a form-fill-seal package.

Optical Properties

Important optical properties in plastic packaging materials are:

Light transmission—the ratio of the intensity of a light source measured with the film interposed to the intensity without the film. It gives no indication of image distortion or blurring.

Clarity—the degree of distortion of an object seen through the film.

Haze—a measure of milkiness caused by light scattering by surface imperfections or film inhomogeneities such as crystallites, voids, cross-linked materials, and undissolved additives. Haze obscures visibility for product inspection.

Gloss—measures specular reflection, or the reflectance of light as a mirror reflects. This parameter indicates the ability to produce a sharp image of any light source, giving rise to a pleasing sparkle of the film.

Electrical Properties

Electrical properties can be important, as for the dissipation of static charge in the operating room. This was previously of greater concern when ether was used more widely as an anesthetic and poured from a bottle, resulting in a potential fire hazard. More importantly, static electricity is a hazard to electronic equipment and devices. In addition, dirt and dust are attracted by static to the surface and increase the chance of contamination.

PHYSICOCHEMICAL PROPERTIES

Mass Transfer

Many pharmaceutical preparations must be protected adequately from oxygen, water vapor, carbon dioxide, and other permeants. An effervescent tablet requires a barrier to moisture, for example, whereas an oil-based product must be protected from oxygen-induced oxidation. Unlike glass, plastics are permeable. Barrier properties indicate permeability to water vapor, oxygen, carbon dioxide, etc. In addition, components of the product can permeate through the package. Examples include stabilizing agents such as parabens or antibacterials, flavorants, water vapor, and oils.

Permeation through a plastic barrier depends on the composition of the plastic, permeation area, thickness of the barrier, partial pressure differential of the permeant across the barrier, and time. Fick's law of diffusion describes these phenomena mathematically. Permeation through a plastic also can be affected greatly by additives and the crystalline structure of the plastic. Specific additives, primarily plasticizers, can increase the permeation rate greatly. Highly crystalline plastics such as polypropylene generally exhibit low water-permeation rates. An increase in the size (eg, diameter, molar volume) of a penetrant in a series of chemically similar penetrants generally leads to an increase in solubility and a decrease in diffusion coefficient. Since the permeability coefficient is related to the product of these, its variation with penetrant size is often much less.[2]

As a guide, the approximate relative permeation rates for water vapor, oxygen, and carbon dioxide through the more commonly used plastics in packaging are given in Table 54-1.[3] More extensive compilations of permeation rates for a variety of migrating molecules can be found in the *Polymer Handbook*.[4] The total ingress of gas into a package can be divided into contributions from the separate components, for example, permeation through the lid, bottle, and outer protective overpouch and gross leakage through microscopic cracks and pinholes. This analysis can be performed kinetically to verify container integrity or to resolve manufacturing problems.[5]

Chemical Attack

Resistance to acids, alkalies, fats, solvents, water, and light are important if compatibility with these materials is required. Some plastics are incompatible with plasticizers used with PVC polymers, lipid emulsions, detergents, or antiseptic solutions. Iodine-containing liquids permanently stain many polyolefin compounds after a brief exposure. Absorption of the migrating

Table 54-1. Permeability Rates of Selective Plastic Packaging Materials

	G/100 IN^2/24 HR/MIL @ 37.8°C	CC/100 IN^2/MIL/24 HR/ATM @ 25°C	
PLASTIC	WATER VAPOR	OXYGEN	CARBON DIOXIDE
Nylon			
Type 6	16–22	2.6	10–12
Type 12	4	34–92	153–336
Polyethylene terephthalate	1.0–1.3	3.0–6.0	15–25
Polyethylene			
Low density	1.0–1.5	500	2700
Medium density	0.7	250–535	1000–2500
High density	0.3	185	580
Polypropylene	0.7	150–240	500–800
Polystyrene	7–10	250–350	900
Vinyl			
Nonplasticized	2–5	4–30	4–30
Plasticized	15–40	600	20–500
Vinyl chloride-acetate copolymer			
Nonplasticized	4	15–20	40–70
Plasticized	5–8	20–150	70–800
Polyvinylidene chloride	0.2–0.6	0.8–6.9	3.8–44
Polycarbonate	11	300	1075

From *Modern Plastics Encyclopedia,* vol 64. New York: McGraw-Hill, 1987, p. 554.

chemical forces the polymer chains apart, swells the plastic, and causes stress cracking. This can occur, as well, from solvents used to solvent-bond plastic components.

Rubber exposed to ozone, as from electrostatic dust precipitators, will lose elasticity and become brittle. In this case, chemical reaction of the ozone with the polymer backbone is responsible. Another failure mode involves simply the leaching of components, such as silicone lubricant from rubber syringe plungers, into the contained solution. This increases the particulate burden and can form a visual haze. In some instances pigmentation used in the plastic is attacked chemically and leached by the product.

In the case of plastics used in direct contact with a product—in either dry or liquid form—the length of time that the medication and the container are in contact may determine if problems such as discoloration, leaching, and absorption or adsorption of a constituent of the product may arise. It is possible that both the product and the package containing it could change significantly from the time of manufacture. Lack of visual indication of a reaction at the onset of a stability study does not imply that the reaction(s) was not occurring during the early stages of storage.

In certain instances, a specific set of storage parameters must exist before a reaction is initiated. For many drugs, generally the higher the temperature and humidity in the storage area, the more rapid the chemical attack. For many IV solutions in flexible plastic containers, however, shelf life is limited by water-vapor loss, which is diminished in the presence of high humidity. Other factors that may affect the plastic packaging and product are pH, surface treatment of the plastic, container configuration, type of polymer used, method of package preparation, light transmission, and means of assembly or sterilization.[6]

Theory and experiment have been developed sufficiently to permit prediction of the maximal accumulation of leachables in solution without waiting for the results of shelf-life stability studies. This expedites product development and addresses material/solution compatibility concerns. Accumulation of a leachable material from a container into solution can be limited by any of four physical factors[7]:

1. The initial amount of leachable material present in the container material (total available pool, TAP)
2. The solubility limit of the leachable material in the solution phase
3. The equilibrium partitioning of the leachable component between the container and the solution
4. The rate of migration of the leachable component from the container into solution

The TAP, solubility limit, and equilibrium partitioning can be evaluated for each identified leachable substance. These characteristics then can be used to identify the process that would limit the accumulation of leachable chemicals. The lowest value found determines the limiting accumulation and identifies the limiting mechanism. The solubility limit, equilibrium partitioning, and rate of migration may restrict actual solution accumulation below the total available pool estimate. Kinetic control produces the minimum accumulation estimate, since no matter how fast the rate of migration, a leachable component cannot accumulate in excess of what is thermodynamically available. As an example, the equilibrium solution accumulation of a leachable material, C_e, is given by

$$C_e = (TAP \times W_c)/[(W_c \times E_b) + V_s]$$

where TAP is the μg leachable/g of film, W_c is the weight of the container (in grams), V_s is the solution volume (in liters), and E_b is the equilibrium partitioning constant, the ratio of the concentration of solute in the film to that in water at equilibrium. This can be calculated from the more familiar, and referenced, solvent-solvent partition coefficients. This methodology also can be used to predict the extent of the reverse process, that of adsorption of solution components (drugs or antimicrobial agents) into the plastic.[8]

The concept of solid/liquid partition coefficients is discussed in Chapter 33, *Chromatography*. Similarly, liquid/liquid partition coefficients are discussed in Chapter 36, *Separation*. Additional consideration of failure modes for pharmaceutical applications may be found in Chapter 52, *Stability of Pharmaceutical Products*.

Safety Testing

Numerous testing procedures must be followed to ensure the safety of use of any plastic. Among these are biological, chemical, physical, and pharmacological assessments. A greater degree of safety testing is warranted as the extent of contact of the material with the body increases. Thus, an IV solution container is studied in greater depth than is secondary packaging. Medical devices that are left intact in the human body for prolonged periods of time (vascular grafts, cartilage replacements, pacemakers, or prosthetics) are studied most extensively. Their reactivity and degree of safety and toxicity must be determined. In all cases, it is imperative that the plastic and its processing procedure provide a nonreactive and nontoxic end product.

The official compendia provide procedures for performing certain biological and physicochemical tests on plastic containers; for details, see the United States Pharmacopeia (USP). The principles of these tests are described in the following sections.

BIOLOGIC TESTING PROCEDURES—USP General Chapter <87> and <88> Biological Reactivity determine the suitability of plastic materials intended for use in fabricating containers or accessories for both parenteral and ophthalmic preparations. The procedures for the former determine the reaction of living animal tissues and normal animals to implanted portions of the plastic or injected extracts prepared from it. Depending on the use of the plastic, other biological tests may be performed, such as pyrogenicity, blood compatibility, antigenicity, suitability for use in cardiovascular devices, gene-toxicity reaction, and tissue-toxicity testing.

PHYSIOCHEMICAL TESTING PROCEDURES—Many chemical and physical tests are applied to plastics, the particular ones used depending on the intended applications of the substances. USP General Chapter <661> Physicochemical Testing specifies the physical and chemical properties of plastics used as containers, based on tests with extracts prepared by heating samples with Water for Injection at 70° for 24 hr. Portions of the extract are used to determine Nonvolatile Residue, Residue on Ignition, Heavy Metals, and Buffering Capacity or Reaction, official limits for each of which are specified. Also described is a procedure for determining the light transmission of plastics, with limits for maximum transmission. USP General Chapter <381> Elastomeric Closures and <661> and <671> Moisture Permeation are relevant for the final container/closure system fabricated from the materials. Drug contacting materials must also meet the chemical standards embodied in 21 CFR Part 175 and Part 177 Indirect Food Additives Section. Additionally, the Food and Drug Administration (FDA) has published a guidance document, "Container Closure Systems for Packaging Human Drugs and Biologics," dated May 1999.

The actual product/package should be evaluated under simulated use conditions, including shipping and storage, to ensure product integrity throughout its shelf life. Potential incompatibilities between the primary plastic container and secondary packaging should be addressed to anticipate adulteration of the product. Prolonged exposure to ultraviolet light has been shown to enhance the migration of certain additives that in turn can accelerate the aging characteristics of the plastic and decrease the shelf life of the product. In some instances, incompatibilities that might occur readily can be detected visually; in others, sophisticated extraction techniques must be followed to ascertain the effects storage conditions may have had. For this reason, well-planned stability studies need to be established.

Desirable features used for health care packaging are transparency, thermal stability, physical strength, formability, sealability, biological barrier, radiation resistance, and disposability. Usually one cannot find all the desired properties in a single material, but two or more plastics can be combined into a composite packaging material.

Failure Mode Analysis

After development and subsequent distribution of plastic-packaged items, functional problems may occur occasionally. Resolution of these problems requires analysis of the causative-failure mode. This involves problem isolation, segregating the problem material to a particular batch, for example, to identify potential causative factors. The failed parts are subjected to mechanical, microscopic, and chemical analysis for further determination of how they differ from acceptable parts. The analytical techniques chosen are dictated by the observed mode of failure.

Physical tests, such as mechanical, electrical, and optical determinations, can be performed quickly, and control values exist in the form of manufacturers' specifications, which are readily available. As the problem becomes more precisely focused, more specific and often elaborate testing is performed to isolate the cause further. For example, reduced stiffness of a part may be attributable to lowered molecular weight of the plastic. Microscopic analysis is rapid, and a skilled analyst often can identify the problem as a pinhole, improper seal, delamination of a composite material, or foreign material acting as a stress fracture initiator.

Chemical analysis of impurities that may cause bloom or prevent seal formation is often time consuming because of the tiny amounts present, the large variety of potential compounds, and a lack of control information from the supplier. The expense and variety of chemical instrumentation available requires judicious selection of the approach to be used.

CLASSIFICATION

There are over 100 different polymer types available for use that can be classified further into two subcategories. These are identified as thermoplastics and thermosets (thermosetting plastics). Thermoplastics consist of those plastics that normally are rigid at operating temperatures but can be remelted and reprocessed. Thermosets consist of those plastics that, when subjected to heat, normally will become infusible or insoluble, and as such cannot be remelted.

ADDITIVES AND MODIFIERS

Thermoplastics can be modified greatly and have their properties enhanced by the addition of specific additives. As chemicals may act synergistically, any two safe additives may have the potential to produce undesirable effects when combined. For these reasons, the FDA requires that these blends or combinations be evaluated totally, prior to marketing in product form. Chemical, pharmacological, and biological tests should be conducted to establish safety. Problems involving additives include migration to the surface of molded parts and leaching into aqueous solutions. Additives used routinely in thermoplastic formulations are discussed below.

Lubricants are used to assist processing of the plastic during the molding or extrusion operation, facilitating flow in contact with metal surfaces. A commonly used lubricant in the case of polyethylene is zinc stearate. The quantities employed vary from formulation to formulation.

Stabilizers are used to retard or prevent degradation of the polymer by heat and light during manufacturing as well as to improve its aging characteristics. Common stabilizer families include organometallic compounds, fatty acid salts, and inorganic oxides.

Antioxidants are a special type of stabilizer used to retard oxidation, by inhibiting formation of free radicals. Examples are aromatic amines, hindered phenolics, thioesters, and phosphites. Combinations of antioxidants with other additives may result in undesirable chemical reactions. Recent technology permits introduction of a desiccant directly into plastic packaging for protection of moisture-sensitive products. The process involves entraining a desiccant in the polymer stream for molding into a container wall. It is intended for medical diagnostic and test-strip kits, effervescent drugs, and nutritional products.[9] Some work also has been done with adding the antioxidant vitamin E to plastic packaging,[10] permitting food to taste fresh for longer times. Oxygen scavengers can be incorporated directly into films, closure liners, and container walls and customized to the required oxygen-absorbing capacity.[11]

Plasticizers are used to achieve softness, flexibility, and melt flow during processing. They are used commonly in plastic materials such as vinyls, cellulosics, and propionates. The most common are high boiling liquids, usually phthalates, of which dioctyl phthalate is the most popular.

Antistatic agents are used to prevent the buildup of static charges on the plastic surface.

Slip agents are added primarily to polyolefins (polyethylene and polypropylene) to reduce the coefficient of friction of the material. These particular chemicals result in antitack and antiblock characteristics in the end product.

Dyes and pigments are added to impart color.

Surface treatments of film, by corona discharge or deposition of thin layers of other plastics, improve such properties as ink adherence, adherence to other films, heal sealability, or gas barrier.

PROCESSING

Besides the addition of additives, the manner in which a plastic is formed into the desired configuration can affect the end properties. It is important that process parameters, such as temperature, pressure, and time, be controlled rigidly to ensure batch-to-batch uniformity for plastic objects. If process parameters are not controlled adequately, such deleterious effects on plastic properties as thermal degradation, piece-part stresses, and incorrect physical dimensions may result. Process thermal degradation of a plastic can affect the leaching characteristics of the plastic object, its permeation characteristics, and its long-term stability during the shelf life of the pharmaceutical product. Piece-part stresses may be relieved when the pharmaceutical package is subjected to certain environmental conditions, resulting in package failure during the shelf life of the product. Small stress fractures in the flange of thermoformed trays, introduced during the thermoforming process, for example, may compromise sterility.

The more common plastic-processing methods employed for pharmaceutical packaging components follow.

Injection Molding

Injection molding is an intermittent process, the plastic being heated to a melted or viscous state and then forced into a cavity (mold) at high pressure. The melted material cools in the cavity and solidifies. The mold is then opened and the part removed. A wide range of thermoplastic and several thermosetting materials can be injection molded. Besides threads on bottle caps, very intricate configurations can be obtained by injection molding of plastics.

Extrusion

Extrusion is a continuous process, the plastic being heated to a melted or viscous state and forced under pressure through a die, resulting in a configuration of desired shape. A slit-shaped

die will result in a plastic sheet, and a circular die will yield a tube of plastic. The extruded profile is cooled to a solid state, generally by spraying with or immersion in water, or by using chilled rolls for film material. A wide range of thermoplastic materials can be extruded. Typical extruded profiles used by the pharmaceutical industry are packaging films and medical tubing. Plastic packaging film also is formed by blow extrusion, an extruded tube being blown into a large cylinder and then slit after cooling.

Besides simply imparting a new shape to the molten plastic, the manufacturing process can preferentially orient the molecular chains in a given direction, by stretching the plastic. This in turn affects physical properties such as clarity and impact strength, as the chains are oriented along the load-bearing direction. Crystallites can be formed and oriented to yield increases in strength, albeit at reduction in elongation at break. Barrier properties are improved for polypropylene. Biaxially oriented film has balanced properties if the same extent of stretching is used in each direction. In cast film, orientation in the machine direction is achieved by feeding the film through a series of rolls running at gradually increasing speeds. Rolls are heated sufficiently to bring the film to suitable temperature below the melting point. Transverse orientation is obtained by use of a tenter frame, which has two divergent endless belts fitted with clips. These grip the film, so that as it travels forward, it is drawn transversely at the required draw ratio. Uniaxial orientation is used for high-performance tape.

Composite Film Manufacture

Multilayer plastic structures permit incorporation of disparate properties not otherwise obtainable from one material. These include tailoring of gas barrier, heat sealability, strength, and adhesion to other materials such as paper. They are made by the basic methods of coating, lamination, and coextrusion. Coatings are applied to films as dispersions, as solutions in organic solvents, or as molten material. In metallization, coatings of aluminum are applied by vaporization of the molten metal under vacuum and condense on moving film. Lamination is the most versatile process, permitting joining together of paper and foil, in addition to thermoplastics. Preformed dissimilar films are joined together with heat and adhesives, such as vinyl acetate or polyurethanes. Coextrusion is less expensive than lamination, where applicable, forming the composite structure without separately creating the component webs. From multiple extruders, separate streams of different molten polymers are simultaneously fed to a die that joins them while preventing their intermixing.

Blow Molding

The plastic is heated to a melted or viscous state and formed into a hollow cylinder (parison) either by extrusion or injection molding. If extruded, the parison is cut to the required length and transferred to the blowing cavity (mold). The bottom of the parison is pinched off by the mold, and air is blown into the parison, expanding the viscous plastic to the walls of the cavity, thus forming the desired container shape. The melted material cools in the cavity and solidifies. The mold is opened, and the container removed. Pharmaceutical bottles are blow molded from a wide range of thermoplastic materials, among which polyethylene and polypropylene predominate.

Solvent Casting

A liquid suspension of rubber is deposited on an endless belt, and the solvent is vaporized. The belt carries the rubber material through a heat cabinet to cure it, whereupon the film is stripped off the belt, cooled, and wound onto reels.

Compression Molding

Compression molding is used for thermosetting materials and is an intermittent process. The thermosetting material (powder or a tablet preform) is placed into a heated cavity (mold). The material melts and flows to fill the cavity. The mold is held under pressure until the thermosetting material cures, after which the mold is opened and the part removed. As with injection molding, very intricate configurations can be obtained by compression molding of thermosetting materials.

TYPES AND USES

The following types of plastics are used commonly in healthcare practice; several of their properties and end uses are indicated.

Thermoplastics

The following are used commonly in injection molding, blow molding, extrusion, and fabricated sheeting.

POLYETHYLENE—This polymer, PE, has the molecular structure $(CH_2)_n$ and is the most pervasively used because it affords essential properties for the least cost. The properties of polyethylene vary according to molecular weight and type: low-density (LDPE) or branched, and high-density (HDPE) or linear. The length and number of side-chain branches determine the degree of crystallinity and density. The linear type has a more regular molecular structure, hence is more crystalline, and therefore is stronger, stiffer, more heat resistant, less permeable to gases, and more resistant to oils than LDPE. Additionally, as crystallinity and density increase, opacity, tensile strength, surface hardness, and chemical resistance increase. Silicone oil and surfactants, however, can act as stress-crack agents, leading to crack formation in stressed areas, as the permeants spread apart the polymeric chains.

Both LDPE and HDPE have relatively low water absorption, excellent electrical resistance, and high resistance to most solvents and chemicals and are tasteless and odorless. PE is thus well suited to many applications in which only moderate-to-low heat exposure will be encountered. Its use ranges from containers for liquid or dry products to both laminated and unsupported films for sterile-device packaging and to molded parts for a variety of devices and equipment. Unsupported polyethylene is used for shrink wrapping, stretch wrapping, skin packaging, and bags.

With more tightly packed molecules HDPE has better moisture-barrier properties with less elongation (better tensile strength) than LDPE. It is used widely, when rigidity and barrier properties are preferred, for bottles of solid dosage form products. However, LDPE is used when flexibility is required, for squeeze bottles of sprays and drops, as well as drum liners for bulk solid drugs. Blown films of LDPE have very low haze and high gloss, whereas HDPE films have higher haze, because of crystal-induced light scattering, and are semiglossy. The less crystalline LDPE has a lower melting point with broader melting range than does HDPE and, therefore, is easier to heat seal. The low melting point, however, negates steam sterilization for LDPE, unlike HDPE.

Polyethylene is used as a primary packing film, but its use as a sealant, through the application of heat and pressure, is more important. This application requires strong seals to be made at low temperature that have good hot tack (ie, to maintain seal integrity as the temperature cools). For this purpose, linear low density polyethylene (LLDPE)is used. This resin has reduced side chain branching and combines the clarity and density of LDPE with the toughness of HDPE. These characteristics arise from the molecular structure, resulting from the reaction of HDPE with unsaturated comonomers such as butene, hexene, or octene. The incompatibility of these two types of polymers inhibits the sealant layer from forming a complete

bond, by reducing the number of available bonding sites and thereby reducing the interfacial adhesion. On the other hand, by narrowing the molecular weight range of LLDPE, one can produce film with the same strength at a lower gauge, thus saving cost.[12]

ETHYLENE-VINYL ACETATE (EVA)—Addition of vinyl acetate comonomer to ethylene reduces polymer crystallinity, improving clarity, low-temperature flexibility and toughness, impact strength, and stress-crack and flex-crack resistance and reducing hardness. Melting and heat-seal temperatures are lowered, as are the barrier properties. Increased vinyl acetate concentration also increases polarity, resulting in increased tackiness and adhesion to a variety of substrates. The copolymer also can be cross-linked (chemical bonds form between the polymer chains) by either radiation or addition of organic peroxides. This increases the melting temperature, permitting autoclaving as a sterilization option. Adding vinyl acetate softens the material, resulting in a smoother surface. The copolymer EVA is used in tip protectors, where flex resistance is required, and for low-temperature IV bags.

The two main characteristics controlled in the copolymerization of vinyl acetate and ethylene are crystallinity and molecular weight. Molecular weight is controlled by the addition of radical chain-transfer agents. As the molecular weight of EVA increases, so does the melt viscosity, heat-seal strength, toughness, flexibility, stress-crack resistance, and hot-tack strength. One of the leachables is acetic acid, resulting from the hydrolysis of the acetate esters.

POLYPROPYLENE—Propylene (PP) is clearer than HDPE, and it is stronger, stiffer, and more heat resistant than LDPE. This material is available as the highly crystalline, isotactic polypropylene and the higher-impact grades of atactic and syndiotactic types. *Isotactic* refers to a plastic with the organic groups (R) on the same side of the polymer chain. *Syndiotactic* refers to the alternation of organic groups above and below the polymer chain, and *atactic* signifies no regular sequences of the groups.

Polypropylene is used widely for packaging of solid dosage products. Injection-molded bottles, for example, can be made either with separate lids or with integrally molded lids, which exhibit high flexural strength. Compared to PE, PP offers better resistance to oils, odors, and less tendency to absorb antimicrobial agents agents from bactericidal solutions.

Polypropylenes can be modified with polyethylene or rubber to improve their impact resistance. Higher levels of ethylene lower stiffness and improve clarity. Biaxial orientation also will improve its clarity and mechanical properties, but is difficult to heat seal. It is, however, the nonoriented cast copolymer that is most used for health care packaging. Devices made of this material can be sterilized with steam and ethylene oxide but not radiation, unless modified polypropylenes are used.

The low density polypropylene offers an economic advantage, as more molds can be made from a given weight of the material. Nucleating agents may be added to speed the rate of crystallization, thus shortening the molding cycle, resulting in more economical manufacturing processes and cheaper products.

Because polypropylene is largely chemically resistant, it cannot be solvent bonded. It can, nevertheless, be heat bonded. Bonding by use of adhesives requires surface pretreatment using corona, plasma or flame, or chemical etching. It can be made heat sealable by applying a coating of polyvinylidene chloride or ethylene-polypropylene copolymer.

CYCLIC OLEFIN COPOLYMERS—Cyclic Olefin Copolymers (COCs) represent a new resin family. The resulting films for blister packs combine a high moisture barrier with the easy forming and sealing properties of polyolefins such as PE. A coextruded PP/COC structure offers a gas and water vapor transmission rate equivalent to polyvinylidene dichloride without the halogens.[13]

POLYVINYL CHLORIDE—Polyvinyl chloride (PVC) commonly called vinyl, is next to HDPE the most widely used plastic for drug packaging, largely because of clarity, low cost, and great fabrication flexibility. The term vinyl comes from the monomer structural group (CH_2═CH—), which has many derivatives, such as vinyl chloride (CH_2═CHCl), vinyl acetate (CH_2═$COCOCH_3$), and vinylidene chloride (CH_2═CCl_2, Saran). With this group of vinyl compounds, many polymers are made either as homopolymers of themselves or as copolymers with other vinyl derivatives or other monomeric materials. For example, polyvinylidene chloride (PVDC) resins are, for the most part, copolymers of vinylidene chloride with vinyl chloride, acrylonitrile, and acrylate esters. These are used primarily where high barrier properties to moisture, oxygen, and other chemicals are required.

The versatile vinyl plastics are used to prepare materials ranging from soft, flexible sheeting to rigid, hard tubing. The great variety of PVC resins, with their wide range of physical properties, led to the development of many applications of this material in the fields of pharmacy and medicine. It is used in the manufacture of blood bags, examination gloves, IV solution containers, and pump tubing. An unplasticized form is used in the fabrication of rigid parts for devices. Because unplasticized PVC has glass-like clarity, is inexpensive, and has excellent thermoformability, it makes an appealing blister pack, where it has a dominant market position for the plastic component. It finds limited use in packaging devices because it turns brown when exposed to radiation sterilization and is too heat sensitive for steam sterilization, and degassing ethylene oxide is too lengthy. However, more than 25% of all plastic-based medical devices used in hospitals are made of PVC, because of its weldability, cost, response to heat and pressure, and versatility.[14]

PVC is used in clear bottles rather than HDPE for reasons of better clarity, gloss, better odor barrier, or absorption of fewer flavor components. HDPE, however, can be autoclaved and does not require the extensive additive package of PVC, involving antioxidants, etc.[15]

Flexible PVC has excellent impact and flex-crack resistance at room temperature. As the temperature is lowered, the material becomes stiffer, resulting in decreased flex-crack resistance and impact strength. The type and amount of plasticizer determines the temperature at which the failure mode changes from ductile to a brittle failure. For flexible medical applications such as IV bags and tubing, the plasticizer DEHP (di(2-ethylhexyl)phthalate) is used most often. Because it can leach into solution, the safety of DEHP has been studied extensively throughout the years. No long-term exposure problems have been identified.[16]

Cyclohexanone can be used to bond PVC to most materials. When bonded to DEHP-noncompatible materials, such as polycarbonate and impact-grade polystyrene, a barrier adhesive must be used.

POLYSTYRENE (PS)—This polymer is one of the oldest and most widely used plastics. Its clarity, stiffness, thermoformability and cheap cost are responsible for its use in manufacture of pharmaceutical bottles and tubes, which do not require a gas barrier.

PS has relatively low heat resistance and is attacked by a number of chemical agents, such as phthalate plasticizers in vinyl polymers, resulting in crazing (microcracks). It is available in a clear crystal grade and an increasingly popular rubber-modified impact-resistant grade, in which polystyrene is copolymerized with acrylonitrile and butadiene. The crystal versions craze during most ethylene oxide cycles, but impact grades withstand both gas and radiation sterilization. Polystyrene cannot, however, be autoclaved. While this polymer is inexpensive, the lack of impact strength in the conventional grade and poor optical properties in impact-modified grades limit its use in more demanding applications. Use in drug packaging is also declining because of its poor gas barrier and solvent resistance.

TYVEK (*DUPONT*)—This is a nonwoven, spun-bonded polyethylene that appears white, smooth, and water repellent and offers high tear strength as well as good porosity for sterilization. It is the preferred material for lidding of trays. How-

ever, it is expensive and has poor print quality, and its web varies in thickness and density. It can be used for autoclaving up to 137°. Prior to thermal disintegration, it will become translucent, indicating that its properties have been compromised. A new, nonwoven polypropylene, Securon, withstands steam sterilization over 153°.[17]

IONOMER—Ionomer is used as an inner ply in laminates, offering good heat sealing (even when the seal area is contaminated by liquid or powder) over a wide temperature range for LDPE and oriented PP. Heat sealing usually can proceed faster than by using alternate materials. Ionomers are clear, semiflexible, tough materials with good abrasion resistance, all of which are features valued in sachet and pouch packs. Their expense limits application to those areas such as seal integrity or enhanced puncture resistance where the additional cost can be justified.

Chemically, ionomers are the sodium or zinc salts of ethylene/methacrylic acid polymers. The ionic cross-links occur randomly along the long-chain polymer molecules to produce solid-state properties usually associated with polymers of high molecular weight. Heating ionomers to normal thermoplastic-processing temperatures, however, diminishes these ionic forces, allowing the material to be melt processed in conventional molding and extrusion equipment. The long-chain, semicrystalline hydrocarbon polymer imparts polyolefinic character, chemical inertness, thermal stability, and low water-vapor transmission.

FLUORINE-CONTAINING POLYMERS—Fluoropolymer-Aclar Film (polymonochlorotrifluoroethylene, PCTFE) has extremely low transmission of moisture, is transparent, and can be heat sealed, laminated, printed, thermoformed, metallized, and sterilized. Because it is the most expensive plastic used in the pharmaceutical industry, it is employed only where the most demanding barrier properties are required. Laminated Aclar/PVC sheet is used widely in thermoformed blister packs for moisture-sensitive solid dosage forms.

POLYTETRAFLUOROETHYLENE—Polytetrafluoroethylene (PTFE) or Teflon offers exceptional chemical resistance, compelling its use as a liner for rubber stoppers to protect the package contents from adulteration by stopper components.

POLYURETHANE FOAMS—Polyurethane foams are formed by polymerization in the presence of a foaming agent, which evolves carbon dioxide, and have been used as a replacement for cotton wool in tablet containers. The polyurethane is however light sensitive, thus limiting application to opaque containers or tinted to hide light-catalyzed discoloration.

NYLONS—Nylon is the generic designation for a class of polyamides containing repeating amide groups (—CONH—) connected to methylene units (—CH_2—) in the structure of the polymer. They are characterized by good chemical resistance to most solvents and chemicals, with the exception of strong solutions of certain mineral acids, phenolic compounds, and strong oxidizers. Nylons can be used in the fabrication of precision parts and adapters for devices and equipment. Aerosol valves, for example, have a low wear requirement that is satisfied by nylon's low friction-bearing surfaces. Nylon also is used in the manufacture of certain high end packaging films and laminates, providing clarity and imparting excellent resistance to puncture and abrasion. Because of high cost, poor moisture barrier properties, and poor sterilization survival (it wrinkles during autoclaving and degrades upon irradiation) its success in form/fill/seal food-packaging applications has not made an impact on health-care packaging.

POLYETHYLENE TEREPHTHALATE (PET)—PET is prepared from ethylene glycol and either terephthalic acid or the dimethyl ester of terephthalic acid. Its chemical structure is p-$HO(COC_6H_4COOCH_2CH_2O)_nH$. PET exists in an amorphous state, an oriented and partially crystalline state, and a highly crystalline state. Most applications require orientation and/or crystallization to take advantage of the dramatically increased strength and improved serviceability at high temperatures that result. PET polymers offer many advantages to the container and packaging field. Among those are its high strength, excellent clarity, low transmission rate to gas and water vapor, and sterilizability by all major modes. PET bottles are used for a wide variety of foods and beverages, as well as pharmaceutical containers. Use of PET and glycol modified PET (PETG) for liquid oral dosage form containers is described in detail in the USP.[18] Heavier gauge, semirigid, unoriented polyester is used in the manufacture of blister packs.

POLYCARBONATES—These are formed by condensation of polyphenols such as bisphenol-A with phosgene. The polymers are transparent thermoplastics (although opacifiers are added for some applications), with high strength and high temperature resistance. Because they are expensive, their use is limited to specialty applications where dimensional stability or high-impact resistance are valued, such as in rigid, transparent, blood oxygenator housings. The polycarbonates have hardness properties similar to those of metals and are being used to replace metals in numerous industrial applications. Their use is increasing, partly because of their ability to withstand radiation sterilization.

Creep-resistance is good over a broad range of temperatures, and parts can be molded consistently to tolerances of 0.002 inch/inch. They can be heat or solvent sealed, facilitating fabrication procedures, but this advantage also renders them susceptible to phthalate crazing, when placed in contact with plasticized vinyls.

ACRYLICS—This class includes the polymethacrylates, polyacrylates, and copolymers of acrylonitrile. There are many variations in this class, mainly concerned with the combinations of methacrylate and acrylate esters, as well as acrylonitrile. These plastics are characterized by clarity and unusual optical properties, low water absorption, good electrical resistivity, excellent weatherability, and fair tensile strength. Their heat resistance is low, and care should be taken to keep them below temperatures of 200°F, at which they tend to soften. Acrylics find considerable use in a multiplicity of devices employed in today's hospitals and clinics. A specific application is in the adapters used in solution-administration sets and blood-collection sets.

CELLULOSICS—To be used as a thermoplastic without charring, cellulose must be modified. The range of modification available permits a wide variety of physical characteristics, including toughness, surface gloss, good clarity, good scuff resistance, and high gas permeability. To achieve these properties, the cellulosic alcohol groups are esterified with acetate, butyrate, and/or propionate. Butyrate and propionate are chosen over acetate for applications requiring low-temperature impact strength and dimensional stability. Extruded butyrate and propionate sheeting have good gage uniformity, surface quality, brilliance, and visual effects. Propionate is selected over butyrate and acetate when increases in hardness, tensile strength, and stiffness are important. Increased plasticizer level lowers hardness, stiffness, and tensile strength but increases impact strength. Combined esters such as cellulose acetate propionate and cellulose acetate butyrate are especially popular for medical applications. This family of cellulosics is used in articles such as tubing and special trays for urological or spinal procedures, membranes in dialyzers and some filters, and IV buret housings.

Thermosets

The following are some of the commonly used compression-molded, thermosetting compounds. These plastics are used when good dimensional and temperature stability are required. Parts are fabricated by means of compression-molding techniques. The formaldehyde plastics are obtained by condensation reactions between formaldehyde and substances such as melamine, phenol, and urea.

As a family, the formaldehydes have been found to be of most use in the pharmaceutical industry as closures for glass and/or

plastic containers. By virtue of high resistance to heat, they are used in specific applications where the molded part requires sterilization by steam.

Elastomeric polymers are characterized by high stretchability. This characteristic arises from an extensive, highly crosslinked, thee dimensional structure of the polymer. The more cross links, the stronger and stiffer the product. A greater frequency of unsaturated bonds in the polymer affords greater elasticity, but also poorer resistance to water and oil. The resulting mechanical properties of compressibility and resealability are desirable for parenteral container closures. Compressibility permits sealing of small irregularities in mating surfaces and reclosability affords improved sterility control following puncture after a hypodermic needle has been withdrawn. Butyl and chlorobutyl rubber are used primarily for these applications because of their additional feature of resistance to permeation by oxygen and water vapor. The addition of natural rubber is added to the formulation when better coring resistance to multiple needle penetration is desired.

Following their molding, the stoppers may be glazed by chorination or siliconized to reduce their coefficient of friction. To minimize chemical interaction with container contents, a teflon coating may also be applied.

MELAMINE FORMALDEHYDE—This family of plastics exhibits good-to-excellent dimensional stability. When used in the manufacture of closures, high torque strength and good impact strength are obtained. These plastics also exhibit good resistance to oils, grease, and many organic solvents.

PHENOL FORMALDEHYDE—This type of plastic provides good scratch-resistant parts. It exhibits very low shrinkage and low water-absorption properties. It is, however, a relatively brittle plastic.

UREA FORMALDEHYDE—This plastic exhibits good dimensional stability as well as good strength properties. Articles produced from this material are highly rigid and provide good resistance to alcohols, oils, grease, and some of the weaker acids. These properties permit use for injection-molded heads for collapsible tubes used to contain liquid-based topical products.

APPLICATIONS

Composite materials, incorporating several components or plies, are used often to obtain the numerous advantages of multiple materials, all of which are unavailable from just one component. A stable material forming the bulk of the film is selected, such as PET, which is very popular for flexible packaging, providing dimensional and thermal stability. To this can be added protective coatings, such as barrier materials affording protection from oxygen, water vapor, and gasses. Also available are sealant layers permitting the package to be heat sealed and bonding layers to accommodate printing inks and to bond the various layers together in multiple-ply extrusions or laminations.

HEALTH CARE DEVICE PACKAGING—This is designed to protect medical devices during sterilization and shipping. The material porosity required for steam or ethylene oxide gas sterilization must be considered in conjunction with the need for maintaining a bacterial barrier following sterilization. Some candidate materials must be rejected because they cannot survive the sterilization mode. For example, PVC, unless specially stabilized, turns brown when subjected to radiation sterilization. Polypropylene becomes brittle only months following radiation exposure.

A satisfactory vent bag consists of a porous Tyvek pouch incorporated into a 3-mil or thicker LDPE bag. This permits rapid in- and out-gassing of ethylene oxide, minimizing expensive sterilization and hold-storage times. The thickness represents a compromise between cost and performance, because thinner bags tend to tear, thus occasioning repacking and resterilizing.

For products requiring better protection than that afforded by a flexible pouch, tray packages can be used. Thermoformed trays are the dominant form of sterile packaging, popular because of strict infection-control standards. These may be either preformed or formed on-line. The latter uses thermoform/fill/seal machinery that first unwinds a web of flexible material from a reel, and then heats a section of it, forming it into a container. This is then filled with product and sealed on-line in one continuous operation. Inexpensive PVC or the higher-barrier PETG copolyester often is considered for the blister tray, because of thermoforming capability, appearance, toughness, and dimensional stability. Denesting agents as additives are critical for thermoformed trays, which require materials characterized by high gloss and a high coefficient of friction.[19]

The blister tray subsequently is sealed to a Tyvek or paper lid. Either lidding material is a sterile barrier and permits steam to penetrate the package. Paper may yellow and embrittle, however, during autoclaving. Furthermore, paper is hygroscopic and changes dimension in response to changes in humidity. This can lead to seal failure, as the lidding can pull away from the tray.

Clamshell packaging also affords stronger, infection-resistant containers. Transparency of both blister and clamshell packages allows the user to inspect the contents visually prior to breaking the seal, thus eliminating waste created from opening the wrong package.[20]

Heat-seal coating technology is important to ensure a reliably sterile product. Modified PE often is used for the heat-seal coating. Sealant properties of PE can be modified, depending upon the product requirements, by branching the polymer chain, which decreases its crystallinity and hence density. By decreasing density, the sealing range, elongation, stress-flex resistance, elasticity, and impact strength increase. As density increases, the following properties increase: sealing temperature, tensile strength, stiffness, hardness, barrier properties, and chemical resistance.

The seal occurs as the melt zones of the plastic are forged together to allow the polymer chains to cross the interface and form a bond. Package leaks can arise form poor uniformity of the heat-source application, and bubble and pore formation from foaming of moisture due to improperly dried plastics. Additionally, poor control of mechanical pressure applied during the melting process can result in squeezing the molten plastic out of the melt zone. Hermeticity is measured by dye-penetrant, bubble-emission, pressure decay, microbial-ingress, radioactive, and mass spectrometer systems.[21]

For packaging products high in alcohol content, EMA (ethylene methacrylic acid) copolymers may be used because of their ability to seal despite contact with organic contaminants in the seal area. For bonding to metal foils, PE must be made more hydrophilic. This is accomplished by copolymerizing hydrophobic ethylene with the more hydrophilic EAA (ethylene acrylic acid), EMA, or ionomer. Polypropylene also can be modified by the incorporation of random ethylene monomers in the polymer chain. This confers rubbery character to the sealant, increasing impact strength and flexibility.

Reduction of costs and increased requirements for validation drive technology trends in the packaging area. The future will see reduced material being used or a shift to less expensive materials where possible. The following are examples of these trends: shift from semi-rigid to flexible packaging for IV catheters; thick flexible to thin flexible packaging for syringes; nylon to polyolefin for dressings; coated products to uncoated products; and clean peel to fiber tear for syringes and needles.[22]

BLISTER PACKAGING—This type, for solid dosage forms, uses tray technology similar to that described above, except that the compartments are smaller. It involves forming a heat-softened plastic film into or around a deep-drawn, pocketed mold to make a plastic tray (thermoforming), filling with a solid dosage-form product and sealing with push-through or peelable covering. The forming film, covering, and product must flow at the right rates without sticking. Appropriate heat and pressure must be applied to ensure that permanent sealing will be formed that will protect the product

throughout its shelf life. A critical property of the seal is hot tack, resistance of creep while warm, because as the package is ejected from the heat-seal jig, the still-warm bond line must support its entire weight.[23] It comprises two components: the melt strength of the seal layer at the temperature of the seal and the interfacial adhesion of the sealant layer to the opposite web.

Choice of film thickness affects both material costs and barrier properties. Other considerations are machineability, production rates, depth of the blister, wall thickness and uniformity of the blister, and sealing properties. Unplasticized, or rigid, PVC is the most common material for forming film because it is thermoformed easily, has glass-like clarity, is inexpensive, has high flexural strength, good chemical resistance, low permeability to fats, oils, easy tintability, and has barrier properties that are adequate for many drugs. The typical film thickness of 250 μm (10 mil) can be increased by applying a 25 to 50 μm coating of PVDC (polyvinylidene chloride) that increases the water vapor barrier properties 5- to 10-fold.

For better protection, films are made from PVC and CTFE (chlorotrifluoroethylene, Aclar). Such films are 15-fold less permeable to moisture than is PVC of comparable thickness. Maximal protection from water vapor is provided by biaxially oriented polyamide/aluminum/PVC (nylon-Al-PVC) that gives barrier properties that are immeasurably low. Aluminum makes the material more recyclable. The cost is comparable to that of PVDC-coated PVC. Other materials such as PP, PS, or PET have been tried for blister packs but have not achieved commercial success because of technical difficulties, poor barrier properties, or economic issues. The highest degree of protection is afforded by an all-foil package, which is cold-formable. The aluminum layer consists of several very thin layers rather than a single thick one to ensure that pinholes do not go all the way through the foil.[24] Nondestructive blister inspection devices are used to check for leaks in a 30 s vacuum test cycle.

For blister lidding, the selection of material structures depends on product fragility and whether the blister must be child resistant (CR). Polyester/foil is selected for CR applications. For non-CR applications, lidding material is usually aluminum/paper for fragile products or preprinted aluminum for sturdy products.[25] A standard 25-μm thickness of aluminum is considered to be pinhole free and represents an optimum combination of cost and product protection. The hardness of the aluminum can be optimized either for facilitating a push-through opening or hindering it, if a child-proof feature is desired. Lidding material also is perforated along the sealed seams to prevent it from being peeled from the formed film in one piece. The lidding material has a printing primer on one side and a heat-sealing lacquer on the other, which faces the product and forming film. A value-added feature is a peel-off-push-through foil, offered by a paper-polyester-aluminum laminate. The paper/PET laminate first is peeled from the aluminum, and then the tablet is pushed through the aluminum.[26]

Strip and sachet packaging are other unit-type packs used for tablets, capsules, powders, etc. Multidose packs for solid dosage forms can be made from PS, PVC, polyester, PP, or HDPE. The latter two are preferred for their better barrier properties toward moisture. All can be made child resistant and tamper evident/resistant using innovative closure systems reflecting the versatility of plastic materials. To discourage children from biting a package, a non-toxic bittering agent can be added to the paper side of the blister. To combat counterfeiting, 2D and 3D holographic security paper can be incorporated into blister laminate structures.[27] A paperboard-based sleeve-blister card combines compliance assistance with child resistance, tamper evidence, and elder friendliness. A die-cut hinge in the outer sleeve releases a folded paperboard encased film/foil blister slide card, which can be printed with dosage instructions to aid compliance.[28]

COLLAPSIBLE TUBES AND FLEXIBLE POUCHES— These are used to contain viscous and liquid-based topical products. They usually are constructed of metal or metal-lined, low-density polyethylene, or a laminated material. Tubes are fabricated by rolling and heat-sealing flat stock into a continuous tube, then trimming to length and attaching the head by injection molding. Metal tubes are airtight, light-proof, and impermeable and offer superior protection. Plastic tubes are lightweight, leakproof, and relatively nonbreakable. In contrast to collapsible metal tubes that flatten as the product is removed, plastic tubes have memory that permits them to retain their original shape after squeezing. Laminate squeeze tubes offer the advantages of plastic with barrier properties close to those of metal. For some applications, an internal liner is used, shielding the product from the seam of the tube, which can cause crystallization.[29]

INTRAVENOUS (IV) SOLUTIONS—Compared with glass bottles, plastic packaging offers nonbreakability and light weight, affording easier transport and handling. Additionally, flexible packaging permits collapsibility, which provides greater protection from aerial contamination. Also, squeezing the bag with a pressure cuff enables rapid administration of large fluid quantities in emergency situations. This puts a burst-strength requirement on both the material and the quality of the seals. A container designed to keep products separated before mixing features a frangible seal between two or more pouch compartments that keeps multiple injectable solutions apart until just before use. [30]

Because of its transparency, durability, autoclavability, and manufacturability at an economical cost, PVC has been a material of choice. The realities of shipping require pinhole resistance. This is offered by flexible, high-yield strength materials like plasticized vinyl, rather than stiffer, more brittle materials like unmodified polyolefins. The polar nature of PVC permits rapid radiofrequency sealing of the bag, incorporating the port tubes for an IV administration set and medication sites. A polyolefin overwrap is used as a water-vapor barrier to prevent excessive moisture loss through the plasticized PVC.

Automatic packaging of IV solutions can be accomplished with an aseptic form (blow molding)-fill-seal system for rigid containers or a seal-fill-seal system for flexible containers. The latter requires preformed plastic film, which is reel-fed onto a forming manifold and side-sealed to form a tube, which is then filled. After incorporating fitments and closures, the final seal is made, and the completed container is cut from the web of material. The materials used are primarily polyethylene, polypropylene, and polyolefin, modified with rubber to increase yield strength for flexible containers. Composite materials may be used, incorporating a heat-seal layer facing the solution, an economical bulk layer for strength, and a polyester outside layer for scuff resistance and glossy appearance.[31] Additional considerations for IV container materials may be found in the chapters on parenteral preparations and intravenous admixtures.

PHARMACEUTICAL COIL—Coil material is placed into bottles of solid oral dosage forms to prevent damage during shipping and handling. These materials include cotton, rayon, polyester, or an HDPE plastic spring. Purified rayon filler is a fibrous form of bleached, regenerated cellulose. Purified polyester filler contains a number of additives (eg, antistatic, antiabrasives, lubricants) and therefore could leach residues. But it has the advantage that it does not contain water, as do the other hydrophilic polymeric materials.[32]

COATING MEDICAL DEVICES—Medical devices that are exposed to body fluids are often coated for protection against corrosion or for lubricity to surfaces. In one process, Parylene C, the low molecular weight dimer of para-chloro-xylylene is vapor deposited under vacuum on the device part, where it immediately polymerizes by a free radical process. Typical anticorrosion applications include blood pressure sensors, cardiac-assist devices, prosthetic components, bone pins, electronic circuits, ultrasonic transducers, bone-growth stimulators, and brain probes. Applications to promote lubricity include mandrels, injections needles, cannulae, and catheters.[33]

STERILIZATION

For plastic medical devices and packaging materials, a number of sterilizing agents have been used, including (1) steam, (2) gas, and (3) irradiation (cobalt and electron discharge). Of these agents, steam can be used on only a few polymers because of their inability to withstand heat without distortion. The following commonly used plastic types generally can withstand steam sterilization at temperatures of 121° C: polypropylene, high-density polyethylene, polycarbonate, PVC for certain applications, and all thermosets.

The most commonly used procedure for sterilizing plastic devices is gas sterilization. Some of the gases available are (1) 100% ethylene oxide, (2) 88%/12% mixtures of Freon and ethylene oxide, and (3) 80%/20% or 90%/10% mixtures of carbon dioxide and ethylene oxide.

Gas sterilization cannot be used for containers of aqueous products because side-reaction products such as ethylene glycol and 2-chloroethanol are formed. Ethylene oxide itself is carcinogenic. It also can react with body proteins and certain material leachables to form immunogenic compounds that can elicit hypersensitivity reactions. For this reason, regulatory permissible limits have been established for residual levels of ethylene oxide. To meet these limits, packaged products are degassed prior to shipping or use. Degassing properties depend upon geometry, heat history, storage conditions, contact with other plastics, and type of secondary packages used. Because of this complexity, degassing hold times must be determined for each product.

Newer gas sterilization technologies also have been developed. These include vaporized hydrogen peroxide, plasma processes such as Plazlyte and Sterrad, and chloride dioxide as well as PureBright, an intense, pulsed-light process. These modalities may afford a wider availability of materials compatible with these processes.[34]

Irradiation can cause degradation or cross-linking of certain polymers. This is particularly serious for polypropylene. Although a radiation-stable form of PP has been developed, it may not be suitable for multiple sterilizations.[35] PVC loses hydrochloric acid upon irradiation, decomposing into unstable fragments, which may then cross-link. This dehydrochlorination leads to the formation of conjugated double bonds, which impart yellow discoloration to the plastic. As part of the additive package to make PVC more radiation resistant, blue dyes are added to mask the yellow coloration. Radical-chain terminators also are added to minimize chain scission. Plastic packaging films, based on the total amount of radiolysis products, may be ranked in order of decreasing stability as polystyrene, polyester, PTFE, nylon, PVDC, PC, PP, HDPE, and LDPE.[36] As cobalt 60 becomes depleted, it may require increased exposure times to produce the same dose level. These longer exposure times can increase oxidative degradation on the packaging.[37] Certain polymers like polyethylene acquire improved tensile and impact strength because of the cross-linking attendant with radiation. The effect upon composite materials may not necessarily correlate with the properties of the individual components. Thus, the loss of strength of a cellulosic film may not be noticed if the film is supported by polyethylene or foil. The suitability of packaging materials subjected to various sterilization methods is discussed further in the chapter on sterilization.

QUALITY-CONTROL CONSIDERATIONS

The selection and approval of a polymer type (and a specific compound within that type) is as important as the need to check it routinely against the criteria used in its selection. The following basic areas of control and/or procedures are recommended regarding an ongoing quality-control program.

Tissue-cell toxicity testing (or a similar toxicity test) should be conducted to provide assurance that the material being used is nontoxic or falls within the toxicity range originally specified.

Characterization analysis should be conducted to provide assurances that the proper polymer type is used and that the physical parameters have not been altered, which in turn could affect the function of the product/package. Such techniques as infrared spectrometric analysis, density, melt-flow, and thermal and rheological tests can assist in providing the necessary assurances.

Any plastic part or package should be inspected routinely on an incoming basis for dimensional and attribute variables against statistically accepted sampling plans such as MIL-STD-105D.

ENVIRONMENTAL CONSIDERATIONS

Disposal is a critical issue, as the volume of solid waste continues to increase and the capacity of landfill sites dwindles. Hospitals are coming under increasing pressure as communities frown upon incineration and disposal costs escalate. Of the total municipal waste generated, plastic packaging accounts for only 4% by weight. While paper accounts for 50% and glass about 25%, plastics draw much of the concern of environmentalists, because of their persistence (nondegradability) in landfill sites. Additionally, plastics are increasingly displacing conventional packaging materials, and on a volume basis, bulky and resilient plastic bottles constitute more of a problem than their weight percentage would imply. The problem is being addressed from a number of standpoints.

Disposal is a complex issue, involving both economics and regulatory requirements. Often the plastic selection alternatives depend upon many factors, such as the mode of disposal or incineration versus landfill. For example, PVC has come under attack because it forms hydrochloric acid when incinerated, necessitating expensive scrubbing systems to neutralize the acid. Dioxins also may be formed if the incineration system is not optimized. If incineration will not be used to dispose of the medical waste for a given location, however, these objections become irrelevant for the particular case.

A global trend to eliminate solvents and volatile organic compounds (VOCs) is leading to interest in packaging with UV-curable inks. The faster drying times associated with this technology helps increase throughput.[38]

In response to their customers, hospital supply manufacturers are reducing the amount of packaging material accompanying their products. Some are working with hospitals to establish successful recycling efforts. This requires convenience of collection, viable reprocessing technology, markets for waste-derived products, and good economics. The individual plastic resins must be sorted prior to being reprocessed for relatively undemanding, nonpackaging applications such as fiberfill. Under some circumstances, homogeneous resins, such as the PET in beverage bottles, can be recycled more easily than composite materials, because of this sorting issue. Plastics manufacturers can, however, incorporate scrap into one of the component layers of some composite materials, making such items potentially recyclable. Recycling of PVC infusion containers is hampered by the difficulties involved in separating metal and rubber components, disinfecting, and drying the products to render them suitable for processing.[14] Nevertheless, the industry is investing in sorting technology and reclamation capacity to create commercially viable recycling programs.[39]

There is a trend toward elimination of folding carton dispensers. These are being replaced with polyethylene bagged shipments, used in conjunction with reusable bins in the hospital central supply rooms and pharmacies, to reduce both cost and waste. In the mail service prescription industry, air bubble mailers are preferred to corrugated ones because of the savings in postage, labor, and incorporation of interior cushioning.

SUMMARY

Before the selection of a plastic for a packaging application is made, all the functional and safety requirements must be specified. These requirements are restated in terms of engineering

and scientific material-testing parameters. Candidate materials are reviewed and selected on the basis of the most economical solution that addresses the critical needs. Within each polymer class, properties may be altered to an extent by modifying molecular weight, copolymerizing with other polymers, or blending in particular additives. Often, composite materials are used to combine the advantages of the individual components. Proper sterilization procedures, including adequate degassing, must be identified to obtain a sterile product that is nontoxic. Once designed, the product/package must demonstrate physical and chemical stability in formal stability studies over the shelf life of the product. An ongoing quality assurance program should be designed to ensure that packaging- product requirements are maintained. After use, disposal of the packaging is becoming more of an issue from economic and environmental standpoints. For more specific and in-depth information, consult the *Bibliography*.

REFERENCES

1. ISO 11607. *Packaging for Terminally Sterilized Medical Devices*.
2. Comyn J, ed. *Polymer Permeability*. New York: Elsevier, 1985.
3. *Modern Plastics Encyclopedia,* vol 64. New York: McGraw-Hill, 1987, p 554.
4. Yasuda H, Stannett V. In: Brandrup J, Immergut EH, eds. *Polymer Handbook,* 2nd ed. New York: Wiley, 1975.
5. Rabinow B, Payton R, Raghavan N. *J Pharm Sci* 1986; 75: 808.
6. Wang YJ, Chien YW. *Sterile Pharmaceutical Packaging: Compatibility and Stability* (Tech Rpt #5). Philadelphia: PDA, 1984.
7. Sanchez IC, Chang SS, Smith LE. *Polymer News* 1980; 6:249.
8. Jenke DR, et al. *Int J Pharm* 1992; 78:115.
9. *Mod Plastics* 1997; (May).
10. *Business Wire* 1996; (Jul 9).
11. Forcinio H. *Pharm Technol* 1999; (Nov):30.
12. *Med Device & Diag Ind* 1998; (Aug).
13. Forcinio H. *Pharm Technol* 2000; (May):26
14. *Med Device Technol* 1991; (Jun).
15. Jenkins WA, Osborn, KR. *Packaging Drugs and Pharmaceuticals*. Lancaster, PA: Technomic Publishing, 1993, p 113.
16. Van Dooren AA. *Pharm Weekbl [Sci]* 1991; 13(3):109.
17. *Pharm Med Pkg News* 1995; (Mar).
18. USP/NF. Rockville, MD: USPC, 661.
19. *Pkg Week* 1997; (Apr 24).
20. Smith RC Jr, ed. *Medical & Healthcare Marketplace Guide,* 12th ed. IDD Enterprises, 1996.
21. *Med Device & Diag Ind* 2000; (Jan):186.
22. *Med Device Technol* 1999; (April):26.
23. Pilchuk R. *Pharm Technol* 2000; (Nov):68.
24. Pilchuk R. *Pharm Technol* 2000; (Nov):68.
25. Forcinio H. *Pharm Technol* 2000; (Jun):24.
26. Reiterer F. *Pharm Technol* 1991; (Mar):74.
27. Forcinio H. *Pharm Technol* 1999; (Nov):30.
28. Forcinio H, *Pharm Technol* 2000; (Jun):24
29. *Pharm Med Pkg News* 1996; (Feb).
30. Forcinio H. *Pharm Technol* 1999; (Jan):28.
31. Lambert P. *Pharm Technol* 1991; (Apr): 48.
32. Taborsky CJ, Mehta U, Kusz M, et al. *Pack Technol* 2000; (Mar):44.
33. *Med Plast and Biomat* 1996; (Mar).
34. *Pkg Week* 1997; n41(Apr 24):17.
35. *Pharm Med Pkg News* 1996; (Sep).
36. *Pkg Technol Eng* 1996; (Jun).
37. Dyke D. *Med Device & Diag Ind* 1996; (Jan).
38. *Pkg Week* 1997; n41(Apr 24):17.
39. *J Vinyl Technol* 1991; 13(2).

BIBLIOGRAPHY

Briston JH. *Plastic Films,* 3rd ed. New York: Wiley, 1988.

Brostow W, Corneliussen RD. *Failure of Plastics*. New York: Macmillan, 1986.

Comyn J, ed. *Polymer Permeability.* New York: Elsevier, 1985.

Crank J, Park GS, eds. *Diffusion in Polymers*. New York: Academic, 1968.

Dean DA. *The Packaging of Pharmaceuticals* (Int Pkg Conf, CONEX 85 (Oct 22–25, 1985), vol 1. Beijing: China Pkg Technol Assoc, 1985, p 287.

Dean DA. *Plastics in Pharmaceutical Packaging*. England: Antony Rowe Ltd, 1990.

Dean DA , Evans ER, Hall IH. *Pharmaceutical Packaging Technology*. New York: Taylor and Francis, 2000.

Finlayson KM. *Plastic Film Technology, High Barrier Plastic Films for Packaging*, vol 1. Lancaster, PA: Technomic Publ Co, 1989.

Jenkins WA, Osborn KR,.*Packaging Drugs and Pharmaceuticals*. Lancaster, PA: Technomic Publishing, 1993.

Modern Plastics Encyclopedia, vol 68. New York: McGraw-Hill, 1992.

Osborn KR, Jenkins WA, *Plastic Films*. Lancaster, PA: Technomic Publishing, 1992.

Yasuda H, Stannett V. Permeability coefficients. In *Polymer Handbook,* 2nd ed, Brandrup J, Immergut EH, eds. New York: Wiley, 1975.

Rodriguez F. *Principles of Polymer Systems,* 2nd ed. New York: Hemisphere, 1982.

Wiley Encyclopedia of Packaging Technology. New York: Wiley, 1986.

CHAPTER 55

Pharmaceutical Necessities

William J Reilly, Jr, RPh, MBA

The practice of pharmacy is an ever-evolving profession that has changes occurring regularly. The costs associated with discovering new compounds are increasing at such a rapid rate that many in healthcare and government don't think is capable of being maintained. The Food and Drug Administration has levied the largest fines in its history over the last couple of years on manufacturers who have failed to comply with what is known as current Good Manufacturing Procedures. The government's approval of new products is in a downward trend, this being a result of companies wanting to provide more information in the filings and the FDA frequently issuing 'not approvable' letters, requiring the sponsor company to conduct additional studies. At the community level, more independent pharmacies are closing their doors or selling their patient lists to national or regional chains. Hospital settings are seeing a greater degree of mergers so that economics are more favorable.

In addition to the profession, pharmacy education has undergone dramatic changes in the United States since the last edition of the *Remington* was published. Now the PharmD degree is the entry-level degree for everyone wanting to practice pharmacy. With this degree, the focus on the clinical aspects of pharmacy has an even greater emphasis on the educational process. Much of this is at the expense of basic pharmaceutics and in some instances, because of course loads, electives such as industrial pharmacy courses are not conveniently taken by students. The pharmaceutical industry used to be able to hire graduates with pharmacy degrees for positions in production, quality control, and dosage-form development because of the breadth of understanding the graduate had of pharmaceutical processes. Unfortunately, gaining this understanding is becoming increasingly difficult unless a student pursues an advanced degree in industrial pharmacy or pharmaceutics.

Regardless of what is happening within the profession, the educational system or the industry, it is imperative that pharmacists in all practice settings know it is their obligation to understand what is used to prepare a medication, whether by commercial means or by extemporaneously compounding it in a practice setting. This chapter does not address the legal aspects of community compounding by a pharmacist, nor does it explain all the specifics of formulating a product for commercial manufacturing. The intent of this information is to inform the practicing pharmacist and other interested individuals in understanding commercial formulations which ingredients are necessary for creating a drug product. These substances, known as excipients, are useful in both the community and commercial settings, although they might be used differently. The excipients described include antioxidants and preservatives, emulsifying and suspending agents, ointment bases, solvents, and miscellaneous ingredients. A more detailed review of these excipients and their commercial applicability to dosage-form development can be found in the *Handbook of Pharmaceutical Excipients* (Rowe, Sheskey, and Weller, eds.) as well as other chapters in this edition of *Remington*.

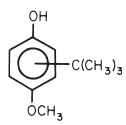
ANTIOXIDANTS AND PRESERVATIVES

An antioxidant is a substance capable of inhibiting oxidation, which may be added for this purpose to pharmaceutical products subject to deterioration by oxidative processes, as for example the development of rancidity in oils and fats or the inactivation of some medicinals in the environment of their dosage forms. A preservative is, in the common pharmaceutical sense, a substance that prevents or inhibits microbial growth, which may be added to pharmaceutical preparations for this purpose to avoid consequent spoilage of the preparations by microorganisms. Both antioxidants and preservatives have many applications in making medicinal products.

ALCOHOL—page 1080.
BENZALKONIUM CHLORIDE—page 1626.
BENZETHONIUM CHLORIDE—page 1627.
BENZYL ALCOHOL—page 1627.

BUTYLATED HYDROXYANISOLE

Phenol, (1,1-dimethylethyl)-4-methoxy-, Tenox BHA



tert-Butyl-4-methoxyphenol [25013-16-5] $C_{11}H_{16}O_2$ (180.25).

Preparation—By an addition interaction of *p*-methoxyphenol and 2-methylpropene. US Pat 2,428,745.

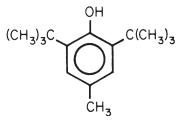
Description—White or slightly yellow, waxy solid having a faint, characteristic odor.

Solubility—Insoluble in water; 1 g in 4 mL alcohol, 2 mL chloroform or 1.2 mL ether.

Uses—An antioxidant in cosmetics and pharmaceuticals containing fats and oils.

BUTYLATED HYDROXYTOLUENE

Phenol, 2,6-bis(1,1-dimethylethyl)-4-methyl-, Butylated Hydroxytoluene Crystalline; Tenox BHT



2,6-Di-*tert*-butyl-*p*-cresol [128-37-0] $C_{15}H_{24}O$ (220.35).

Preparation—By an addition interaction of *p*-cresol and 2- methylpropene. US Pat 2,428,745.

Description—White, tasteless crystals with a mild odor; stable in light or air; melts at 70°C.

Solubility—Insoluble in water; 1 g in 4 mL alcohol, 1.1 mL chloroform, or 1.1 mL ether.

Uses—An antioxidant employed to retard oxidative degradation of oils and fats in various cosmetics and pharmaceuticals.

CHLOROBUTANOL

2-Propanol, 1,1,1-trichloro-2-methyl-, Chlorbutol; Chlorbutanol; Acetone Chloroform; Chloretone

$(CCl_3)C(CH_3)_2OH$

1,1,1-Trichloro-2-methyl-2-propanol [57-15-8] $C_4H_7Cl_3O$ (177.46); *hemihydrate* [6001-64-5] (186.46).

Preparation—Chloroform undergoes chemical addition to acetone under the catalytic influence of powdered potassium hydroxide.

Description—Colorless to white crystals of a characteristic, somewhat camphoraceous odor and taste; anhydrous melts about 95°C; hydrous melts about 76°C; boils with some decomposition between 165° and 168°C.

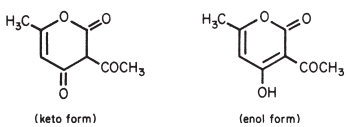
Solubility—1 g in 125 mL water, 1 mL alcohol or about 10 mL glycerin; freely soluble in chloroform, ether, or volatile oils.

Incompatibilities—The anhydrous form must be used to prepare a clear solution in liquid petrolatum. It is decomposed by alkali; ephedrine is sufficiently alkaline to cause its breakdown with the formation of ephedrine hydrochloride, which will separate from a liquid petrolatum solution. It is only slightly soluble in water, hence alcohol must be used to dissolve the required amount in certain vehicles. Trituration with antipyrine, menthol, phenol, and other substances produce a soft mass.

Uses—Topically, as a solution in clove oil as a dental analgesic. It has local anesthetic potency to a mild degree and has been employed as an anesthetic dusting powder (1 to 5%) or ointment (10%). It has antibacterial and germicidal properties. When administered orally, it has much the same therapeutic use as chloral hydrate. Hence, it has been employed as a sedative and hypnotic. It has been taken orally to allay vomiting due to gastritis.

DEHYDROACETIC ACID

Keto form: 2H-Pyran-2,4(3H)-dione, 3-acetyl-6-methyl-,



Enol form: 3-acetyl-4-hydroxy-6-methyl-2*H*-pyran-2-one [520-45-6] (keto), [771-03-9] (enol) $C_8H_8O_4$ (168.15).

Preparation—By fractional distillation of a mixture of ethyl acetoacetate and sodium bicarbonate, maintaining almost total reflux conditions, allowing only ethanol to be removed. The residue is distilled under vacuum. *Org Syn Coll III:* 231, 1955.

Description—White to creamy-white crystalline powder melting about 110°C with sublimation.

Solubility—1 g dissolves in 25 g acetone, 18 g benzene, 5 g methanol, or 3 g alcohol.

Uses—Preservative.

ETHYLENEDIAMINE

1,2-Ethanediamine

$H_2NCH_2CH_2NH_2$

Ethylenediamine [107-15-3] $C_2H_8N_2$ (60.10).

Caution—Use care in handling because of its caustic nature and the irritating properties of its vapor.

Note—It is strongly alkaline and may readily absorb carbon dioxide from the air to form a nonvolatile carbonate. Protect it against undue exposure to the atmosphere.

Preparation—By reacting ethylene dichloride with ammonia, then adding NaOH and distilling.

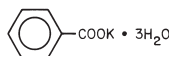
Description—Clear, colorless, or only slightly yellow liquid, with an ammonia-like odor and strong alkaline reaction; miscible with water and alcohol; anhydrous boils 116 to 117°C and solidifies at about 8°C; volatile with steam; a strong base and readily combines with acids to form salts with the evolution of substantial heat.

Uses—A pharmaceutical necessity for Aminophylline Injection. It is irritating to skin and mucous membranes. It also may cause sensitization characterized by asthma and allergic dermatitis.

ETHYL VANILLIN—page 1064.
GLYCERIN—pages 1081 and 1423.
HYPOPHOSPHOROUS ACID—page 1086.
PHENOL—page 1087.
PHENYLETHYL ALCOHOL page 1066.
PHENYLMERCURIC NITRATE—see RPS-19, page 1270.

POTASSIUM BENZOATE

Benzoic Acid, Potassium Salt



[582-25-2] $C_7H_5KO_2$ (160.21) (anhydrous).

Description—Crystalline powder.

Solubility—Soluble in water or alcohol.

Uses—Preservative.

POTASSIUM METABISULFITE

Dipotassium Pyrosulfite

[16731-55-8] $K_2S_2O_5$ (222.31).

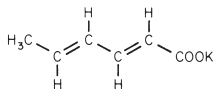
Description—White crystals or crystalline powder with an odor of SO_2. Oxidizes in air to the sulfate. May ignite on powdering in a mortar if too much heat develops.

Solubility—Freely soluble in water; insoluble in alcohol.

Uses—Antioxidant.

POTASSIUM SORBATE

2,4-Hexadienoic Acid, *(E,E)-*, Potassium Salt; 2,4-Hexadienoic Acid, Potassium Salt; Potassium 2,4-Hexadienoate



Potassium *(E,E)*-sorbate; potassium sorbate [590-00-1] [24634-61-5] $C_6H_7KO_2$ (150.22).

Preparation—Sorbic acid is reacted with an equimolar portion of KOH. The resulting potassium sorbate may be crystallized from aqueous ethanol. US Pat 3,173,948.

Description—White crystals or powder with a characteristic odor; melts about 270°C with decomposition.

Solubility—1 g in 4.5 mL water, 35 mL alcohol, >1000 mL chloroform, or >1000 mL ether.

Uses—A water-soluble salt of sorbic acid used in pharmaceuticals to inhibit the growth of molds and yeast. Its toxicity is low, but it may irritate the skin.

SASSAFRAS OIL—page 1069.
SODIUM BENZOATE—see RPS-19, page 1271.

SODIUM BISULFITE

Sulfurous acid, monosodium salt; Sodium Hydrogen Sulfite; Sodium Acid Sulfite; Leucogen

Monosodium sulfite [7631-90-5] $NaHSO_3$ and sodium metabisulfite ($Na_2S_2O_5$) in varying proportions; yields 58.5 to 67.4% of SO_2.

Description—White or yellowish white crystals or granular powder with the odor of sulfur dioxide; unstable in air.

Solubility—1 g in 4 mL water; slightly soluble in alcohol.

Uses—An antioxidant and stabilizing agent. Epinephrine hydrochloride solutions may be stabilized by the addition of small quantities of the salt. It also is used to help solubilize kidney stones. It is useful for removing permanganate stains and for solubilizing certain dyes and other chemicals.

SODIUM METABISULFITE

Disulfurous acid, disodium salt

Disodium pyrosulfite [7681-57-4] $Na_2S_2O_5$ (190.10).

Preparation—Formed when sodium bisulfite undergoes thermal dehydration. It also may be prepared by passing sulfur dioxide over sodium carbonate.

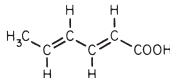
Description—White crystals or white to yellowish crystalline powder with an odor of sulfur dioxide; on exposure to air and moisture, it is slowly oxidized to sulfate.

Solubility—1 g in 2 mL water; slightly soluble in alcohol; freely soluble in glycerin.

Uses—A reducing agent. It is used in easily oxidized pharmaceuticals, such as epinephrine hydrochloride and phenylephrine hydrochloride injections, to retard oxidation.

SORBIC ACID

2,4-Hexadienoic acid, *(E,E)-, 2,4-Hexadienoic acid*



(6)

(E,E)-Sorbic acid; Sorbic acid [22500-92-1] [110-44-1] $C_6H_8O_2$ (112.13).

Preparation—By various processes. Refer to US Pat 2,921,090.

Description—Free-flowing, white, crystalline powder, with a characteristic odor; melts about 133°C.

Solubility—1 g in 1000 mL water, 10 mL alcohol, 15 mL chloroform, 30 mL ether, or 19 mL propylene glycol.

Uses—A mold and yeast inhibitor. It also is used as a fungistatic agent for foods, especially cheeses.

THIMEROSAL—see RPS-19, page 1271.

COLORING, FLAVORING, AND DILUTING AGENTS

The use of properly colored and flavored medicinal substances, although offering no particular therapeutic advantage, is of considerable importance psychologically. A water-clear medicine is not particularly acceptable to most patients and, in general, is thought to be inert. Many very active medicinal substances are quite unpalatable, and the patient may fail to take the medicine simply because the taste or appearance is objectionable. Disagreeable medication can be made both pleasing to the taste and attractive by careful selection of the appropriate coloring, flavoring, and diluting agents. Therefore, judicious use of these substances is important in securing patient cooperation in taking or using the prescribed medication and continued compliance with the prescriber's intent.

Coloring Agents or Colorants

Coloring agents may be defined as compounds employed in pharmacy solely for the purpose of imparting color. They may be classified in various ways (eg, inorganic or organic). For the purpose of this discussion two subdivisions are used: *Natural Coloring Principles* and *Synthetic Coloring Principles*. The members of these groups are used as colors for pharmaceutical preparations, cosmetics, and foods and as bacteriological stains and diagnostic agents.

NATURAL COLORING PRINCIPLES

Natural coloring principles are obtained from mineral, plant, and animal sources. They are used primarily for artistic purposes; as symbolic adornments of natives; as colors for foods, drugs, and cosmetics; and for other psychological effects.

Mineral colors frequently are termed pigments and are used to color lotions, cosmetics, and other preparations, usually for external application. Examples are Red Ferric Oxide and Yellow Ferric Oxide, titanium dioxide, and carbon black.

The term pigment also is applied generically to plant colors by phytochemists. Many plants contain coloring principles that may be extracted and used as colorants (eg, chlorophyll). Anattenes are obtained from annatto seeds and give yellow-to-orange water-soluble dyes. Natural beta-carotene is a yellow color extracted from carrots and used to color margarine. Alizarin is a reddish yellow dye obtained from the madder plant. The indigo plant is the source of a blue pigment called indigo. Flavones, such as riboflavin, rutin, hesperidin, and quercetin, are yellow pigments. Saffron is a glycoside that gives a yellow color to drugs and foods. Cudbear and red saunders are two other dyes obtained from plants. Most plant colors now have been characterized and synthesized, however, and those with the desirable qualities of stability, fastness, and pleasing hue are available commercially as synthetic products.

Animals have been a source of coloring principles from the earliest periods of recorded history. For example, Tyrian purple, once a sign of royalty, was prepared by air oxidation of a colorless secretion obtained from the glands of a snail *(Murex brandaris)*. This dye now is known to be 6,6′-dibromoindigo, and has been synthesized, but cheaper dyes of the same color are available. Cochineal from the insect *Coccus cacti* contains the bright-red coloring principle carminic acid, a derivative of anthraquinone. This dye is no longer used in foods and pharmaceuticals because of *Salmonella* contamination.

SYNTHETIC COLORING PRINCIPLES

Synthetic coloring principles date from 1856 when WH Perkin accidentally discovered mauveine, also known as a Perkin's purple, while engaged in unsuccessful attempts to synthesize quinine. He obtained the dye by oxidizing aniline containing *o*- and *p*-toluidines as impurities. Other discoveries of this kind followed soon after, and a major industry grew up in the field of coal-tar chemistry.

The earliest colors were prepared from aniline, and for many years all coal-tar dyes were called aniline colors, irrespective of their origin. The coal-tar dyes include more than a dozen well-defined groups among which are nitroso-dyes, nitro-dyes, azo-dyes, oxazines, thiazines, pyrazolones, xanthenes, indigoids, anthraquinones, acridines, rosanilines, phthaleins, quinolines, and others. These in turn are classified, according to their method of use, as acid dyes and basic dyes, or direct dyes and mordant dyes.

Certain structural elements in organic molecules, called chromophore groups, give color to the molecules, eg, azo (—N═Noj), nitroso (—N═O), nitro (—NO_2), azoxy (—N═N(O)—), carbonyl (>C═O), and ethylene (>C═C<). Other such combinations augment the chromophore groups, eg, methoxy, hydroxy, and amino groups and are known as auxochromes.

STABILITY—Most dyes are relatively unstable chemicals because of their unsaturated structures. They are subject to fading because of light, metals, heat, microorganisms, oxidizing and reducing agents, plus strong acids and bases. In tablets, fading may appear as spotting and specking.

USES—Most synthetic coloring principles are used in coloring fabrics and for various artistic purposes. They also find application as indicators, bacteriological stains, diagnostic aids, reagents in microscopy, etc.

Many coal-tar dyes originally were used in foodstuffs and beverages without careful selection or discrimination between

those that were harmless and those that were toxic and without any supervision as to purity or freedom from poisonous constituents derived from their manufacture.

After the passage of the Food and Drugs Act in 1906, the US Department of Agriculture established regulations by which a few colors came to be known as permitted colors. Certain of these colors may be used in foods, drugs, and cosmetics, but only after certification by the Food and Drug Administration (FDA) that they meet certain specifications. From this list of permitted colors may be produced, by skillful blending and mixing, other colors that may be used in foods, beverages, and pharmaceutical preparations. Blends of certified dyes must be recertified.

The word permitted is used in a restricted sense. It does not carry with it the right to use colors for purposes of deception, even though they are permitted colors, for all food laws have clauses prohibiting the coloring of foods and beverages in a manner so as to conceal inferiority or to give a false appearance of value.

The certified colors are classified into three groups: FD&C dyes, which legally may be used in foods, drugs, and cosmetics; D&C dyes, which legally may be used in drugs and cosmetics; and external D&C dyes, which legally may be used only in externally applied drugs and cosmetics. There are specific limits for the pure dye, sulfated ash, ether extractives, soluble and insoluble matter, uncombined intermediates, oxides, chlorides, and sulfates. As the use status of these colors is subject to change, the latest regulations of the FDA should be consulted to determine how they may be used—especially since several FD&C dyes formerly widely used have been found to be carcinogenic even when pure and, therefore, have been banned from use.

The Coal-Tar Color Regulations specify that the term externally applied drugs and cosmetics means drugs and cosmetics that are applied only to external parts of the body and not to the lips or any body surface covered by mucous membrane. No certified dye, regardless of its category, legally may be used in any article that is to be applied to the area of the eye.

Lakes are calcium or aluminum salts of certified dyes extended on a substrate of alumina. They are insoluble in water and organic solvents and hence are used to color powders, pharmaceuticals, foods, hard candies, and food packaging.

The application of dyes to pharmaceutical preparations is an art that can be acquired only after an understanding of the characteristics of dyes and knowledge of the composition of the products to be colored has been obtained. Specific rules for the choice or application of dyes to pharmaceutical preparations are difficult to formulate. Each preparation may present unique problems.

Preparations that may be colored include most liquid pharmaceuticals, powders, ointments, and emulsions. Some general hints may be offered in connection with solutions and powders, but desired results usually can be obtained only by a series of trials. In general, an inexperienced operator tends to use a much higher concentration of the dye than is necessary, resulting in a dull color. The amount of dye present in any pharmaceutical preparation should be of a concentration high enough to give the desired color and low enough to prevent toxic reactions and permanent staining of fabrics and tissues.

Liquids (Solutions)—The dye concentration in liquid preparations and solutions usually should come within a range of 0.0005% (1 in 200,000) and 0.001% (1 in 100,000), depending upon the depth of color wanted and the thickness of column to be viewed in the container. With some dyes, concentrations as low as 0.0001% (1 in 1,000,000) may have a distinct tinting effect. Dyes are used most conveniently in the form of stock solutions.

Powders—White powders usually require the incorporation of 0.1% (1 in 1000) of a dye to impart a pastel color. The dyes may be incorporated into the powder by dry-blending in a ball mill or, on a small scale, with a mortar and pestle. The dye is incorporated by trituration and geometric dilution. Powders also may be colored evenly by adding a solution of the dye in alcohol or some other volatile solvent having only a slight solvent action on the powder being colored. When this procedure is employed, the solution is added in portions, with thorough mixing after each addition, after which the solvent is allowed to evaporate from the mixture.

Many of the syrups and elixirs used as flavoring and diluting agents are colored. When such agents are used, no further coloring matter is necessary. The use of colored flavoring agents is discussed in a subsequent section. However, when it is desired to add color to an otherwise colorless mixture, one of the agents described in the first section may be used.

INCOMPATIBILITIES—FD&C dyes are mainly anionic (sodium salts) and hence are incompatible with cationic substances. Since the concentrations of these substances are generally very low, no precipitate is evident. Polyvalent ions such as calcium, magnesium, and aluminum also may form insoluble compounds with dyes. A pH change may cause the color to change. Acids may release the insoluble acid form of the dye.

Flavoring Agents

FLAVOR

The word flavor refers to a mixed sensation of taste, touch, smell, sight, and sound, all of which combine to produce an infinite number of gradations in the perception of a substance. The four primary tastes—sweet, bitter, sour, and saline—appear to result partly from physicochemical and partly from psychological action. Taste buds (Fig 55-1), located mainly on the tongue, contain very sensitive nerve endings that react, in the presence of moisture, with the flavors in the mouth, and as a result of physicochemical activity, electrical impulses are produced and transmitted via the seventh, ninth, and tenth cranial nerves to the areas of the brain that are devoted to the perception of taste. Some of the taste buds are specialized in their function, giving rise to areas on the tongue that are sensitive to only one type of taste. The brain, however, usually perceives taste as a composite sensation, and accordingly, the components of any flavor are not readily discernible. Children have more taste buds than adults and hence are more sensitive to tastes.

Taste partly depends on the ions that are produced in the mouth, but psychologists have demonstrated that sight (color) and sound also play a definite role when certain reflexes become conditioned through custom and association of sense perceptions. Thus, in the classic experiments of Pavlov demonstrating conditioned reflexes, the ringing of a bell or the showing of a circle of light caused the gastric juices of a dog to flow, although no food was placed before it, and much of the enjoyment derived from eating celery is due to its crunchy crispness as the fibrovascular bundles are crushed. The effect of color is just as pronounced; oleomargarine is unpalatable to most people when it is uncolored, but once the dye has been incorporated, gourmets frequently cannot distinguish it from butter. Color and taste must coincide (eg, cherry flavor is associated with a red color).

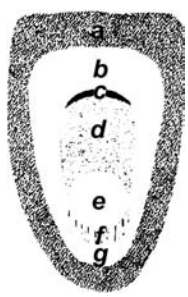


Figure 55-1. Upper surface of the tongue. *a,* Taste receptors for all tastes; *b,* sweet, salty, and sour; *c,* salty and sour; *d,* sour only; *e,* no taste sensation; *f,* sweet and sour; *g,* bitter, sweet, and sour. (Adapted from Crocker EC. *Flavor.* New York: McGraw-Hill, 1945, p 22.)

Persons suffering from a head cold find their food much less palatable than usual because their sense of smell is impaired, and if the nostrils are held closed, raw onions taste sweet, and it is much easier to ingest castor oil and other nauseating medicines. The volatility of a substance is an important factor that is influenced by the warmth and moisture of the mouth, since the more volatile a compound, the more pronounced its odor. The sense of smell detects very minute amounts of material and is usually much more sensitive in detecting the presence of volatile chemicals, but the tongue is able to detect infinitesimal amounts of some vapors if it is protruded from the mouth so that solution of the gases in the saliva may take place. In this manner traces of sulfur dioxide can be detected in the air, since it dissolves in the saliva and creates a sour taste.

Flavors described as hot are those that exert a mild counterirritant effect on the mucosa of the mouth; those that are astringent and pucker the mouth contain tannins and acids that produce this effect by reacting with the lining of the mouth, and wines possess a bouquet because of the odor of the volatile constituents. Indian turnip (Jack-in-the-pulpit) owes its flavor largely to the stinging sensation caused by the minute acicular crystals of calcium oxalate that penetrate the mucous membrane.

Other physiological and physical factors that also may affect taste are coarseness or grittiness due to small particles (eg, ion-exchange resins). Antidiarrheal preparations have a chalky taste. Menthol imparts a cool taste because it affects the coldness receptors. Mannitol gives a cool sensation when it dissolves because its negative heat of solution will cause the temperature to drop. For this reason, mannitol often is used as the base for chewable tablets.

There is a definite threshold of taste for every substance, which varies somewhat with the individual and with the environment. Experienced chefs taste their delicacies at the temperature at which they will be served, since heat and cold alter the flavor of many preparations. Thus, lemon loses its sour taste entirely at an elevated temperature, and other flavors become almost nonvolatile, tasteless, and odorless when cooled sufficiently. In addition to the influence of temperature, the sensitivity of each individual must be considered. For example, it has been determined by experiment that the amount of sugar that can just be detected by the average individual is about 7 mg. However, this amount cannot be tasted by some, and it is definitely sweet to others.

People are more sensitive to odor than to taste. There are about 10,000 to 30,000 identifiable scents, of which the average person can identify about 4000. Women are more sensitive to odors than men. Additional insights can be obtained by reading Beauchamp GK, et al: *Tasting and Smelling,* (New York: Academic, 1997) and Cagan RH, et al: *Neural Mechanisms in Taste* (Boca Raton, FL: CRC Press, 1989).

PRESERVATION OF FLAVORS—Most monographs of official products contain specific directions for storage. Proper methods of storage are essential to prevent deterioration, which in many instances results in destruction of odor and taste. Under adverse conditions undesirable changes occur because of one or a combination of the following: enzymatic activity, oxidation, change in moisture content, absorption of odors, activity of microorganisms, and effects of heat and light. In certain products some of the changes wrought by these factors are desirable, as when esters are formed because of the activity of enzymes and when blending and mellowing results from the interchange of the radicals of esters (transesterification).

One method for protecting readily oxidizable substances, such as lemon oil, from deteriorating, and thus preserving their original delicate flavor, is to microencapsulate them by spray-drying. The capsules containing the flavors then are enclosed in various packaged products (eg, powdered gelatins) or tablets, which are flavored deliciously when the capsule is disintegrated by mixing and warming with water or saliva.

CORRELATION OF CHEMICAL STRUCTURE WITH FLAVOR AND ODOR—The compounds employed as flavors in vehicles vary considerably in their chemical structure, ranging from simple esters (methyl salicylate), alcohols (glycerin), and aldehydes (vanillin) to carbohydrates (honey) and the complex volatile oils (anise oil). Synthetic flavors of almost any desired type are now available. These frequently possess the delicate flavor and aroma of the natural products and also the desirable characteristics of stability, reproducibility, and comparatively low cost. Synthetic products such as cinnamaldehyde and benzaldehyde, first officially recognized when several of the essential oils became scarce during World War II, have been used widely.

There is a close relationship between chemical structure and taste. Solubility, the degree of ionization, and the type of ions produced in the saliva definitely influence the sensation interpreted by the brain.

Sour taste is caused by hydrogen ions, and it is proportional to the hydrogen ion concentration and the lipid solubility of the compound. It is characteristic of acids, tannins, alum, phenols, and lactones. Saltiness is due to simultaneous presence of anions and cations (eg, KBr, NH_4Cl, and sodium salicylate). High-molecular-weight salts may have a bitter taste. Sweet taste is due to polyhydroxy compounds, polyhalogenated aliphatic compounds, and α-amino acids. Amino and amide groups, especially if the positive effect is balanced by the proximity of a negative group, may produce a sweet taste. Sweetness increases with the number of hydroxy groups, possibly because of increased solubility. Imides such as saccharin and sulfamates such as cyclamates are intensely sweet. Cyclamates have been removed from the market because they reportedly cause bladder tumors in rats. Free bases such as alkaloids and amides such as amphetamines give bitter tastes. Polyhydroxy compounds with a molecular weight greater than 300, halogenated substances, and aliphatic thio compounds also may have bitter tastes. Unsaturation frequently bestows a sharp, biting odor and taste on compounds.

No precise relationship between chemical structure and odor has been found. There are no primary odors, and odors blend into each other. Polymerization reduces or destroys odor, high valency gives odor, and unsaturation enhances odor. A tertiary carbon atom often will give a camphoraceous odor, esters and lactones have a fruity odor, and ketones have a pleasant odor. Strong odors often are accompanied by volatility and chemical reactivity.

SELECTION OF FLAVORS

The proper selection of flavors for disguising nauseating medicines aids in their ingestion. Occasionally, sensitive patients have become nauseated sufficiently to vomit at the thought of having to take disagreeable medication, and it is particularly difficult to persuade children to continue to use and retain distasteful preparations. There is a need to know the allergies and idiosyncrasies of the patient; thus, it is foolish to use a chocolate-flavored vehicle for the patient who dislikes the flavor or who is allergic to it, notwithstanding the fact that this flavor is generally acceptable.

FLAVORING METHODOLOGY

Each flavoring problem is unique and requires an individual solution. The problem of flavoring is further complicated because flavor and taste depend on individual preferences. In solving flavoring problems the following techniques have been used:

Blending—Fruit flavors blend with sour taste; bitter tastes can be blended with salty, sweet, and sour tastes; salt reduces sourness and increases sweetness; chemicals such as vanillin, monosodium glutamate, and benzaldehyde are used for blending.

Overshadow—Addition of a flavor whose intensity is longer and stronger than the obvious taste (eg, methyl salicylate, glycyrrhiza, and oleoresins).

Physical—Formation of insoluble compounds of the offending drug (eg, sulfonamides); emulsification of oils; effervescence (eg, magnesium citrate solution); high viscosity of fluids to limit contact of drug with the tongue; and mechanical procedures such as coating tablets are physical methods to reduce flavoring problems.

Chemical—Absorption of the drug on a substrate or formation of a complex of the drug with ion-exchange resins or complexing agents.

Physiological—The taste buds may be anesthetized by menthol or mint flavors.

Flavors, as used by the pharmacist in compounding prescriptions, may be divided into four main categories according to the type of taste that is to be masked, as follows:

Salty Taste—Cinnamon syrup has been found to be the best vehicle for ammonium chloride and other salty drugs such as sodium salicylate and ferric ammonium citrate. In a study of the comparative efficiency of flavoring agents for disguising salty taste, the following additional vehicles were arranged in descending order of usefulness: orange syrup, citric acid syrup, cherry syrup, cocoa syrup, wild cherry syrup, raspberry syrup, glycyrrhiza elixir, aromatic elixir, and glycyrrhiza syrup. The last-named is particularly useful as a vehicle for the salines by virtue of its colloidal properties and the sweetness of both glycyrrhizin and sucrose.

Bitter Taste—Cocoa syrup was found to be the best vehicle for disguising the bitter taste of quinine bisulfate, followed, in descending order of usefulness, by raspberry syrup, cocoa syrup, cherry syrup, cinnamon syrup, compound sarsaparilla syrup, citric acid syrup, licorice syrup, aromatic elixir, orange syrup, and wild cherry syrup.

Acrid or Sour Taste—Raspberry syrup and other fruit syrups are especially efficient in masking the taste of sour substances such as hydrochloric acid. Acacia syrup and other mucilaginous vehicles are best for disguising the acrid taste of substances such as capsicum, since they tend to form a colloidal protective coating over the taste buds of the tongue. Tragacanth, unlike acacia, may be used in an alcoholic vehicle.

Oily Taste—Castor oil may be made palatable by emulsifying with an equal volume of aromatic rhubarb syrup or with compound sarsaparilla syrup. Cod liver oil is disguised effectively by adding wintergreen oil or peppermint oil. Lemon, orange, and anise or combinations of these are also useful. It is better to mix most of the flavor with the oil before emulsifying it, and then the small remaining quantity can be added after the primary emulsion is formed.

Those flavors that are most pleasing to the majority of people are associated with some stimulant of a physical or physiological nature. This may be a CNS stimulant such as caffeine, which is the reason so many enjoy tea and coffee as a beverage, or it may be a counterirritant such as one of the spices that produce a *biting* sensation or an agent that *tickles* the throat such as soda water. Sherry owes its sharp flavor to its acetaldehyde content, and some of the volatile oils contain terpenes that are stimulating to the mucous surfaces.

SELECTION OF VEHICLES

Too few pharmacists realize the unique opportunity they have in acquainting physicians with a knowledge of how to increase both the palatability and efficacy of their prescribed medicines through the judicious selection of vehicles. Because of the training pharmacists receive, their knowledge of the characteristics of various pharmaceuticals and therapeutic agents and their technique and skill in preparing elegant preparations are well developed, so that they are qualified admirably to advise concerning the proper use of vehicles.

A large selection of flavors is available as well as a choice of colors, so that one may prescribe a basic drug for a prolonged period but by changing the vehicle from time to time, the taste and appearance are so altered that the patient does not tire of the prescription or show other psychological reactions to it.

The statement of the late Dr Bernard Fantus that "the best solvent is the best vehicle" helps to explain the proper use of a flavoring vehicle. For example, a substance that is soluble in alcohol (eg, phenobarbital) will not leave an alcoholic vehicle readily to dissolve in the aqueous saliva.

WATERS—These are the simplest of the vehicles and are available with several flavors. They contain no sucrose, a fact to be considered at times, since sucrose under certain circumstances may be undesirable. They are likewise nonalcoholic, another fact that frequently influences vehicle selection.

ELIXIRS—These have added sweetness that waters lack, and they usually contain alcohol, which imparts an added sharpness to the flavor of certain preparations, making the latter more pleasing to the taste. Elixirs are suitable for alcohol-soluble drugs.

SYRUPS—These vehicles, like elixirs, offer a wide selection of flavors and colors from which to choose. Their specific value, however, lies particularly in the fact that they are intensely sweet and contain little or no alcohol, a combination that makes them of singular value as masking agents for water-soluble drugs.

Vehicles consisting of a solution of pleasantly flavored volatile oils in syrup or glycerin (1:500) have been employed successfully in producing uniform and stable preparations. These vehicles are prepared by adding 2 mL of the volatile oil, diluted with 6 mL of alcohol, to 500 mL of glycerin or syrup, which has been warmed gently. The solution is added a little at a time with continuous shaking; then sufficient glycerin or syrup is added to make 1000 mL and mixed well.

Alcohol solutions of volatile oils are sometimes used as stock solutions for flavoring pharmaceuticals.

A listing of substances, most of them official, used as flavors, flavored vehicles, or sweeteners, is given in Table 55-1. Additional information on flavoring ingredients may be obtained in

Table 55-1. Flavoring Agents

Acacia syrup	Lavender oil
Anethole	Lemon oil
Anise oil	Lemon tincture
Aromatic elixir	Mannitol
Benzaldehyde	Methyl salicylate
Benzaldehyde elixir, compound	Nutmeg oil
Caraway	Orange, bitter, elixir
Caraway oil	Orange, bitter, oil
Cardamom oil	Orange flower oil
Cardamom seed	Orange flower water
Cardamom spirit, compound	Orange oil
Cardamom tincture, compound	Orange peel, bitter
Cherry juice	Orange peel, sweet, tincture
Cherry syrup	Orange spirit, compound
Cinnamon	Orange syrup
Cinnamon oil	Peppermint
Cinnamon water	Peppermint oil
Citric acid	Peppermint spirit
Citric acid syrup	Peppermint water
Clove oil	Phenylethyl alcohol
Cocoa	Raspberry juice
Cocoa syrup	Raspberry syrup
Coriander oil	Rosemary oil
Dextrose	Rose oil
Eriodictyon	Rose water
Eriodictyon fluidextract	Rose water, stronger
Eriodictyon syrup, aromatic	Saccharin
Ethyl acetate	Saccharin calcium
Ethyl vanillin	Saccharin sodium
Fennel oil	Sarsaparilla syrup, compound
Ginger	Sorbitol solution
Ginger fluidextract	Spearmint
Ginger oleoresin	Spearmint oil
Glucose	Sucrose
Glycerin	Syrup
Glycyrrhiza	Thyme oil
Glycyrrhiza elixir	Tolu balsam
Glycyrrhiza extract	Tolu balsam syrup
Glycyrrhiza extract, pure	Vanilla
Glycyrrhiza fluidextract	Vanilla tincture
Glycyrrhiza syrup	Vanillin
Honey	Wild cherry syrup
Iso-Alcoholic elixir	

Burdock GA, *Fenaroli's Handbook of Flavor Ingredients,* Cleveland: CRC, 1994.

ACACIA SYRUP—page 1070.

ANISE OIL

Aniseed Oil; Star Anise Oil

The volatile oil distilled with steam from the dried, ripe fruit of *Pimpinella anisum* Linné (Fam *Umbelliferae*) or from the dried, ripe fruit of *Illicium verum* Hooker filius (Fam *Magnoliaceae*).

Note—If solid material has separated, carefully warm the oil until it is completely liquefied, and mix it before using.

Constituents—The official oil varies somewhat in composition, depending upon whether it was obtained from *Pimpinella anisum* or the star anise, *Illicium verum*. Anethole is the chief constituent of both oils, occurring to the extent of 80 to 90%. Methyl chavicol, an isomer of anethole, and anisic ketone [$C_{10}H_{12}O_2$] also are found in both oils, as are small amounts of many other constituents.

Description—Colorless or pale yellow, strongly refractive liquid, having the characteristic odor and taste of anise; specific gravity 0.978 to 0.988; congeals not below 15.

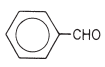
Solubility—Soluble in 3 volumes of 90% alcohol.

Uses—Extensively as a flavoring agent, particularly for licorice candies. It has been given as a carminative in a dose of about 0.1 mL.

AROMATIC ELIXIR—page 1071.

BENZALDEHYDE

Artificial Essential Almond Oil



Benzaldehyde [100-52-7] C_7H_6O (106.12).

Preparation—By the interaction of benzal chloride with lime in the presence of water. Benzal chloride is obtained by treating boiling toluene with chlorine.

Description—Colorless, strongly refractive liquid, with an odor resembling that of bitter almond oil and a burning aromatic taste; affected by light; specific gravity 1.041 to 1.046; boils about 180°C, solidifies about -56.5°C, and on exposure to air it gradually oxidizes to benzoic acid.

Solubility—Dissolves in about 350 volumes of water; miscible with alcohol, ether, chloroform, or fixed and volatile oils.

Uses In place of bitter almond oil for flavoring purposes; it is much safer than the latter because it contains no hydrocyanic acid. It also is used extensively in perfumery and in the manufacture of dyestuffs and many other organic compounds, such as aniline, acetanilid, or mandelic acid.

Compound Benzaldehyde Elixir—Preparation: Dissolve benzaldehyde (0.5 mL) and vanillin (1 g) in alcohol (50 mL); add syrup (400 mL), orange flower water (150 mL), and sufficient purified water, in several portions, shaking the mixture thoroughly after each addition, to make the product measure 1000 mL; then filter, if necessary, until the product is clear. Alcohol Content: 3 to 5%.

Uses—A useful vehicle for administering bromides and other salts, especially when a low alcoholic content is desired.

CARDAMOM SEED

Cardamom Fruit; Cardamom; Ceylon or Malabar Cardamom

The dried ripe seed of *Elettaria cardamomum* (Linné) Maton (Fam. *Zingiberaceae*). It should be removed recently from the capsule.

Constituents—A volatile oil, the yield of which is 1.3% from Malabar Ceylon Seeds and 2.6% from Mysore-Ceylon Seeds. Fixed oil is present to the extent of 10%, also starch, mucilage, etc.

Uses—A flavor. For many years it was employed empirically as a carminative.

Cardamom Oil—The volatile oil distilled from the seed of *Elettaria cardamomum* (Linné) Maton (Fam *Zingiberaceae*). Varieties of the oil contain d-α-terpineol $C_{10}H_{17}OH$, both free and as the acetate; 5 to 10% cineol $C_{10}H_{18}O$; and limonene $C_{10}H_{16}$. The Ceylon Oil, however, contains the alcohol 4-terpineol (4-carbomenthenol) $C_{10}H_{17}OH$, the terpenes, terpinene and sabinene, and acetic and formic acids, probably combined as esters. Description and solubility: Colorless or very pale yellow liquid possessing the aromatic, penetrating, and somewhat camphoraceous odor of cardamom and a persistently pungent, strongly aromatic taste; affected by light; specific gravity 0.917 to 0.947. Miscible with alcohol; dissolves in 5 volumes of 70% alcohol. Uses: A flavor.

CHERRY SYRUP—page 1070.

CINNAMON

Saigon Cinnamon; True Cinnamon; Saigon Cassia

The dried bark of *Cinnamomum loureirii* Nees (Fam. *Lauraceae*). It contains, in each 100 g, not less than 2.5 mL of volatile oil.

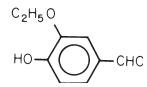
Uses—A flavoring agent. Formerly, it was used as a carminative.

Cinnamon Oil (Cassia Oil; Oil of Chinese Cinnamon)—The volatile oil distilled with steam from the leaves and twigs of *Cinnamomum cassia* (Nees) Nees ex Blume (Fam *Lauraceae*), rectified by distillation; contains not less than 80%, by volume, of the total aldehydes of cinnamon oil. Cinnamaldehyde is the chief constituent. Description and solubility: Yellowish or brownish liquid, becoming darker and thicker on aging or exposure to the air, with the characteristic odor and taste of cassia cinnamon; specific gravity 1.045 to 1.063. Soluble in an equal volume of alcohol, 2 volumes of 70% alcohol, or an equal volume of glacial acetic acid. Uses: A flavor. It formerly was used in a dose of 0.1 mL for flatulent colic.

COCOA SYRUP—page 1070.
CORIANDER—page 1069.
DENATONIUM BENZOATE—page 1085.

ETHYL VANILLIN

Benzaldehyde, 3-ethoxy-4-hydroxy-, Bourbanal; Ethovan; Vanillal; Vanirome



3-Ethoxy-4-hydroxybenzaldehyde [121-32-4] $C_9H_{10}O_3$ (166.18).

Preparation—By reacting *o*-ethoxyphenol with formaldehyde and *p*-nitrosodimethylaniline in the presence of aluminum and water.

Description—Fine, white or slightly yellowish crystals; odor and taste similar to those of vanillin; affected by light; solutions are acid to litmus; melts about 77°C.

Solubility—1 g in about 100 mL water at 50°C; freely soluble in alcohol, chloroform, ether, or solutions of fixed alkali hydroxides.

Uses—A flavor, like vanillin, but stronger.

EUCALYPTUS OIL

The volatile oil distilled with steam from the fresh leaf of *Eucalyptus globulus* Labillardiére or of some other species of *Eucalyptus* L'Heritier (Fam *Myrtaceae*). It contains not less than 70% of $C_{10}H_{18}O$ *(eucalyptol)*.

Constituents—The most important constituent is eucalyptol (cineol). Other compounds include d-α-pinene, globulol, pinocarveol, pinocarvone, and several aldehydes.

Description—Colorless or pale yellow liquid, with a characteristic, aromatic, somewhat camphoraceous odor, and a pungent, spicy, cooling taste; specific gravity 0.905 to 0.925 at 25°C.

Solubility—Soluble in 5 volumes of 70% alcohol.

Uses—A flavoring agent and an expectorant in chronic bronchitis. It also has bacteriostatic properties. This oil may be toxic.

FENNEL OIL

The volatile oil distilled with steam from the dried ripe fruit of *Foeniculum vulgare* Miller (Fam *Umbelliferae*).

Note—If solid material has separated, carefully warm the oil until it is completely liquefied, and mix it before using.

Constituents—Anethole $C_{10}H_{12}O$ is the chief constituent, occurring to the extent of 50 to 60%. Some of the other constituents are d-pinene, phellandrene, dipentene, fenchone, methylchavicol, anisaldehyde and anisic acid.

Description—Colorless or pale yellow liquid, with the characteristic odor and taste of fennel; specific gravity 0.953 to 0.973; congealing temperature is not below 3°C.

Solubility—Soluble in 8 volumes of 80% alcohol or in 1 volume of 90% alcohol.

Uses—A flavoring agent. It formerly was employed in a dose of 0.1 mL as a carminative.

GLYCYRRHIZA

Licorice Root; Liquorice Root; Sweetwood; Italian Juice Root; Spanish Juice Root

The dried rhizome and roots of *Glycyrrhiza glabra* Linné, known in commerce as Spanish Licorice, or of *Glycyrrhiza glabra* Linné var *glandulifera* Waldstein et Kitaibel, known in commerce as Russian Licorice, or of other varieties of *Glycyrrhiza glabra* Linné, yielding a yellow and sweet wood (Fam. *Leguminosae*).

Constituents—This well-known root contains 5 to 7% of the sweet principle glycyrrhizin, or glycyrrhizic acid, which is 50 times as sweet as cane sugar. There also is present an oleoresinous substance to which it's slight acridity is due. If alcohol or an alkali is used as a menstruum for the root and the preparation is not treated to deprive it of acridity, it will have a disagreeable aftertaste. For this reason boiling water is used for its extraction in both the extract and the fluid-extract.

Description—The USP/NF provides descriptions of Unground Spanish and Russian Glycyrrhizas, Histology, and Powdered Glycyrrhiza.

Uses—Valuable in pharmacy chiefly for its sweet flavor, it is one of the most efficient substances known for masking the taste of bitter substances, like quinine. Acids precipitate the glycyrrhizin and should not be added to mixtures in which glycyrrhiza is intended to mask disagreeable taste. Most of the imported licorice is used by tobacco manufacturers to flavor tobacco. It also is used in making candy.

Pure Glycyrrhiza Extract (Pure Licorice Root Extract)—Preparation: Moisten 1000 g of glycyrrhiza, in granular powder, with boiling water, transfer it to a percolator, and percolate with boiling water until the glycyrrhiza is exhausted. Add enough diluted ammonia solution to the percolate to impart a distinctly ammoniacal odor, then boil the liquid under normal atmospheric pressure until it is reduced to a volume of about 1500 mL. Filter the liquid, and immediately evaporate the filtrate until the residue has a pilular consistency. Pure extract of glycyrrhiza differs from the commercial extract in that it is almost completely soluble in aqueous mixtures. The large amount of filler used in the commercial extract to give it firmness renders it unfit to use as a substitute for the pure extract. Description: Black, pilular mass having a characteristic, sweet taste. Uses: A flavoring agent. One of the ingredients in Aromatic Cascara Sagrada Fluidextract.

Glycyrrhiza Fluidextract (Licorice Root Fluidextract); Liquid Extract of Liquorice—Preparation: To 1000 g of coarsely ground glycyrrhiza add about 3000 mL of boiling water, mix and allow to macerate in a suitable, covered percolator for 2 hr. Then allow the percolation to proceed at a rate of 1 to 3 mL/min, gradually adding boiling water until the glycyrrhiza is exhausted. Add enough diluted ammonia solution to the percolate to impart a distinctly ammoniacal odor, then boil the liquid actively under normal atmospheric pressure until it is reduced to a volume of about 1500 mL. Filter the liquid, evaporate the filtrate on a steam bath until the residue measures 750 mL, cool, gradually add 250 mL of alcohol and enough water to make the product measure 1000 mL, and mix. Alcohol Content: 20 to 24%, by volume. Uses: A pleasant flavor for use in syrups and elixirs to be employed as vehicles and correctives.

GLYCYRRHIZA ELIXIR—page 1071.
GLYCYRRHIZA SYRUP—page 1071.
HONEY—page 1092.
HYDRIODIC ACID SYRUP—page 1071.
ISO-ALCOHOLIC ELIXIR—page 1091.

LAVENDER OIL

Lavender Flowers Oil

The volatile oil distilled with steam from the fresh flowering tops of *Lavandula officinalis* Chaix ex Villars (*Lavandula vera* DeCandolle) (Fam *Labiatae*) or produced synthetically. It contains not less than 35% of esters calculated as $C_{12}H_{20}O_2$ (linalyl acetate).

Constituents—It is a product of considerable importance in perfumery. Linalyl acetate is the chief constituent. Cineol appears to be a normal constituent of English oils. Other constituents include amyl alcohol, d-borneol (small amount); geraniol, lavandulol ($C_{10}H_{18}O$); linalool; nerol; acetic, butyric, valeric, and caproic acids (as esters); traces of d-pinene, limonene (in English oils only), and the sesquiterpene caryophyllene; ethyl n-amyl ketone; an aldehyde (probably valeric aldehyde), and coumarin.

Description—Colorless or yellow liquid, with the characteristic odor and taste of lavender flowers; specific gravity 0.875 to 0.888.

Solubility—1 volume in 4 volumes of 70% alcohol.

Uses—Primarily as a perfume. It formerly was used in doses of 0.1 mL as a carminative.

LEMON OIL

The volatile oil obtained by expression, without the aid of heat, from the fresh peel of the fruit of *Citrus limon* (Linné) Burmann filius (Fam *Rutaceae*), with or without the previous separation of the pulp and the peel. The total aldehyde content, calculated as citral ($C_{10}H_{16}O$), is 2.2 to 3.8% for California-type oil, and 3.0 to 5.5% for Italian-type oil.

Note—Do not use oil that has a terebinthine odor.

Constituents—From the standpoint of odor and flavor, the most noteworthy constituent is the aldehyde citral, which is present to the extent of about 4%. About 90% of d-limonene is present; small amounts of l-α-pinene, β-pinene, camphene, β-phellandrene, and γ-terpinene also occur. About 2% of a solid, nonvolatile substance called citroptene, limettin, or lemon-camphor, which is dissolved out of the peel, also is present. In addition, there are traces of several other compounds: α-terpineol; the acetates of linalool and geraniol; citronellal, octyl, and nonyl aldehydes; the sesquiterpenes bisabolene and cadinene, and the ketone methylheptenone.

When fresh, the oil has the fragrant odor of lemons. Because of the instability of the terpenes present, the oil readily undergoes deterioration by oxidation, acquiring a terebinthinate odor.

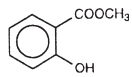
Description—Pale yellow to deep yellow or greenish yellow liquid, with the characteristic odor and taste of the outer part of fresh lemon peel; specific gravity 0.849 to 0.855.

Solubility—In 3 volumes of alcohol; miscible in all proportions with dehydrated alcohol, carbon disulfide, or glacial acetic acid.

Uses—A flavor in pharmaceutical preparations and in certain candies and foods.

METHYL SALICYLATE

Benzoic acid, 2-hydroxy-, methyl ester; Gaultheria Oil; Wintergreen Oil; Betula Oil; Sweet Birch Oil; Teaberry Oil; Artificial Wintergreen Oil; Synthetic Wintergreen Oil



Methyl salicylate [119-36-8] $C_6H_4(OH)COOCH_3$ (152.15); produced synthetically or obtained by maceration and subsequent distillation with steam from the leaves of *Gaultheria procumbens* Linné (Fam *Ericaceae*) or from the bark of *Betula lenta* Linné (Fam *Betulaceae*).

Note—It must be labeled to indicate whether it was made synthetically or distilled from either of the plants mentioned above.

Preparation—Found naturally in gaultheria and betula oils and in many other plants, but the commercial product is usually synthetic, made by esterifying salicylic acid with methyl alcohol in the presence of sulfuric acid, and distilling.

Description—Colorless, yellowish, or reddish liquid, with the characteristic odor and taste of wintergreen; specific gravity (synthetic), 1.180 to 1.185, (from gaultheria or betula), 1.176 to 1.182; boils between 219 and 224°C with some decomposition.

Solubility—Slightly soluble in water; soluble in alcohol or glacial acetic acid.

Uses—A pharmaceutical necessity and counterirritant (local analgesic). As a pharmaceutical necessity, it is used to flavor the official Aromatic Cascara Sagrada Fluidextract and it is equal in every respect to wintergreen oil or sweet birch oil. As a counterirritant, it is applied to the skin in the form of a liniment, ointment, or cream; care should be exercised since salicylate is absorbed through the skin.

Caution—Because it smells like wintergreen candy, it is ingested frequently by children and has caused many fatalities. Keep out of the reach of children.

MONOSODIUM GLUTAMATE

Glutamic acid, monosodium salt, monohydrate

[142-47-2] $C_5H_8NNaO_4H_2O$ (187.13)

Preparation—From the fermentation of beet sugar or molasses or by hydrolysis of vegetable proteins.

Description—White, crystalline powder. The pentahydrate effloresces in air to form the monohydrate.

Solubility—Very soluble in water; sparingly soluble in alcohol.

Uses—Flavoring agent and perfume.

NUTMEG OIL

Myristica Oil NF; East Indian Nutmeg Oil; West Indian Nutmeg Oil

The volatile oil distilled with steam from the dried kernels of the ripe seeds of *Myristica fragrans* Houttuyn (Fam Myristicaceae).

Constituents—It contains about 80% of d-pinene and d-camphene; 8% of dipentene; about 6% of the alcohols d-borneol, geraniol, d-linalool, and terpineol; 4% of myristicin; 0.6% of safrol; 0.3% of myristic acid free and as esters; 0.2% of eugenol and isoeugenol; and traces of the alcohol terpineol-4, a citral-like aldehyde, and several acids, all present as esters.

Description—Colorless or pale yellow liquid with the characteristic odor and taste of nutmeg; specific gravity (East Indian Oil) 0.880 to 0.910, (West Indian Oil) 0.854 to 0.880.

Solubility—In an equal amount of alcohol; 1 volume of East Indian Oil in 3 volumes of 90% alcohol; 1 volume of West Indian Oil in 4 volumes of 90% alcohol.

Uses—Primarily as a flavoring agent. It is used for this purpose in Aromatic Ammonia Spirit. The oil also is employed as a flavor in foods, certain alcoholic beverages, dentifrices, and tobacco; to some extent, it also is used in perfumery. It formerly was used as a carminative and local stimulant to the GI tract in a dose of 0.03 mL. In overdoses, it acts as a narcotic poison. This oil is very difficult to keep and if even slightly terebinthinate is unfit for flavoring purposes.

ORANGE OIL

Sweet Orange Oil

The volatile oil obtained by expression from the fresh peel of the ripe fruit of *Citrus sinensis* (Linné) Osbeck (Fam *Rutaceae*). The total aldehyde content, calculated as decanal ($C_{10}H_{20}O$), is 1.2 to 2.5%.

Note—Do not use oil that has a terebinthine odor.

Constituents—Consists of d-limonene to the extent of at least 90%; in the remaining 5 to 10% are the odorous constituents, among which, in samples of American origin, are n-decylic aldehyde, citral, d-linalool, n-nonyl alcohol, and traces of esters of formic, acetic, caprylic and capric acids.

In addition to most of these compounds, Italian-produced oil contains d-terpineol, terpinolene, α-terpinene, and methyl anthranilate.

Kept under the usual conditions it is very prone to decompose and rapidly acquires a terebinthine odor.

Description—Intensely yellow-orange or deep orange liquid, which possesses the characteristic odor and taste of the outer part of fresh sweet orange peel; specific gravity 0.842 to 0.846.

Solubility—Miscible with dehydrated alcohol or carbon disulfide; dissolves in an equal volume of glacial acetic acid.

Uses—A flavoring agent in elixirs and other preparations.

ORANGE FLOWER WATER—page 1070.

SWEET ORANGE PEEL TINCTURE

Preparation—From sweet orange peel, which is the outer rind of the naturally colored, fresh, ripe fruit of *Citrus sinensis* (Linné) Osbeck (Fam *Rutaceae*), by Process M. Macerate 500 g of the sweet orange peel (exclude the inner, white portion of the rind) in 900 mL of alcohol, and complete the preparation with alcohol to make the product measure 1000 mL. Use talc as the filtering medium.

The white portion of the rind must not be used, as the proportion of oil, which is only in the yellow rind, is reduced, and the bitter principle hesperidin is introduced.

Alcohol Content—62 to 72%.

Uses—A flavor, used in syrups, elixirs, and emulsions. This tincture was introduced to provide a delicate orange flavor direct from the fruit instead of depending upon orange oil, which so frequently is terebinthinate and unfit for use. The tincture keeps well.

COMPOUND ORANGE SPIRIT

Contains, in each 100 mL, 25 to 30 mL of the mixed oils.

Orange Oil	200 mL
Lemon Oil	50 mL
Coriander Oil	20 mL
Anise Oil	5 mL

Alcohol, a sufficient quantity, to make 1000 mL

Mix the oils with sufficient alcohol to make the product measure 1000 mL.

Alcohol Content—65 to 75%.

Uses—A flavor for elixirs. An alcoholic solution of this kind permits the uniform introduction of small proportions of oils and also preserves orange and lemon oils from rapid oxidation. The pharmacist should buy these two oils in small quantities, since the spirit is made most satisfactorily from oils taken from bottles not previously opened. This will ensure that delicacy of flavor that should always be characteristic of elixirs.

ORANGE SYRUP

Syrup of Orange Peel

Contains, in each 100 mL, 450 to 550 mg of citric acid ($C_6H_8O_7$).

Sweet Orange Peel Tincture	50 mL
Citric Acid (anhydrous)	5 g
Talc	15 g
Sucrose	820 g

Purified Water, a sufficient quantity, to make 1000 mL

Triturate the talc with the tincture and citric acid, and gradually add 400 mL of purified water. Then filter, returning the first portions of the filtrate until it becomes clear, and wash the mortar and filter with enough purified water to make the filtrate measure 450 mL. Dissolve the sucrose in this filtrate by agitation, without heating, and add enough purified water to make the product measure 1000 mL. Mix and strain.

Note—Do not use syrup that has a terebinthine odor or taste or shows other indications of deterioration.

Alcohol Content—2 to 5%.

Uses—A pleasant, acidic vehicle.

PEPPERMINT

American Mint; Lamb Mint; Brandy Mint

Consists of the dried leaf and flowering top of *Mentha piperita* Linné (Fam *Labiatae*).

Uses—The source of green color for Peppermint Spirit (see RPS-19 page 902). The odor of fresh peppermint is due to the presence of about 2% of a volatile oil, much of which is lost on drying the leaves in air. It is cultivated widely both in the US and France. It formerly was used as a carminative.

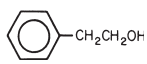
Peppermint Oil—The volatile oil distilled with steam from the fresh overground parts of the flowering plant *Mentha piperita* Linné (Fam *Labiatae*), rectified by distillation and neither partially nor wholly dementholized. It yields not less than 5% of esters, calculated as menthyl acetate $C_{12}H_{22}O_2$, and not less than 50% of total menthol $C_{10}H_{20}O$, free and as esters. Constituents: This is one of the most important of the group of volatile oils. The chief constituent is Menthol (page 1285), which occurs in the levorotatory form; its ester, menthyl acetate, is present in a much smaller amount. Other compounds that are present include the ketone menthone, piperitone, α-pinene, l-limonene, phellandrene, cadinene, menthyl isovalerate, isovaleric aldehyde, acetaldehyde, menthofuran, cineol, an unidentified lactone $C_{10}H_{16}O_2$, and probably amyl acetate.

Description and Solubility—Colorless or pale yellow liquid, with a strong, penetrating odor of peppermint and a pungent taste, followed by a sensation of cold when air is drawn into the mouth; specific gravity 0.896 to 0.908; 1 volume dissolves in 3 volumes of 70% alcohol. Uses: A flavoring agent, carminative, antiseptic, and local anesthetic. It also is used extensively as a flavor in candy, chewing gum, etc.

PEPPERMINT SPIRIT—see RPS-19, page 902.
PEPPERMINT WATER—page 1070.

PHENYLETHYL ALCOHOL

Benzeneethanol; 2-Phenylethanol



Phenethyl alcohol [60-12-8] $C_8H_{10}O$ (122.17); occurs in a number of essential oils such as those of rose, neroli, hyacinth, carnation, and others.

Description—Colorless liquid with a rose-like odor and a sharp, burning taste; solidifies at −27°C; specific gravity 1.017 to 1.020.

Solubility—1 g in 60 mL water or <1 mL alcohol, chloroform, or ether; very soluble in fixed oils, glycerin, or propylene glycol; slightly soluble in mineral oil.

Uses—Introduced for use as an antibacterial agent in ophthalmic solutions, but it is of limited effectiveness.

It is used in flavors, as a soap perfume, and in the preparation of synthetic oils of rose and similar flower oils. It is also a valuable perfume fixative.

ROSE OIL

Otto of Rose; Attar of Rose

The volatile oil distilled with steam from the fresh flowers of *Rosa gallica* Linné, *Rosa damascena* Miller, *Rosa alba* Linné, *Rosa centifolia* Linné, and varieties of these species (Fam *Rosaceae*).

Constituents—From the quantitative standpoint the chief components are the alcohols geraniol ($C_{10}H_{18}O$) and l-citronellol ($C_{10}H_{20}O$). The sesquiterpene alcohols farnesol and nerol occur to the extent of 1% and 5 to 10%, respectively. Together, the four alcohols constitute 70 to 75% of the oil. Phenylethyl alcohol, which constitutes 1% of the oil, is an important odoriferous constituent. Other compounds present are linalool, eugenol, nonyl aldehyde, traces of citral, and two solid hydrocarbons of the paraffin series.

Description and Solubility—A colorless or yellow liquid, which has the characteristic odor and taste of rose; at 25°C, a viscous liquid;

on gradual cooling it changes to a translucent, crystalline mass, which may be liquefied easily by warming; specific gravity 0.848 to 0.863 at 30°C, compared with water at 15°C; 1 mL mixes with 1 mL of chloroform without turbidity; on the addition of 20 mL of 90% alcohol to this solution, the resulting liquid is neutral or acid to moistened litmus paper and deposits a crystalline residue within 5 min on standing at 20°C.

Uses—Principally as a perfume. It is recognized officially for its use as an ingredient in Rose Water Ointment and cosmetics.

STRONGER ROSE WATER

Triple Rose Water

A saturated solution of the odoriferous principles of the flowers of *Rosa centifolia* Linné (Fam *Rosaceae*), prepared by distilling the fresh flowers with water and separating the excess volatile oil from the clear, water portion of the distillate.

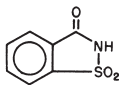
Note—When diluted with an equal volume of purified water, it may be supplied when Rose Water is required.

Description—Nearly colorless and clear liquid that possesses the pleasant odor and taste of fresh rose blossoms; must be free from empyreuma, mustiness, and fungal growths.

Uses—An ingredient in Rose Water Ointment. It sometimes is prepared extemporaneously from concentrates or from rose oil, but such water is not official and rarely compares favorably with the fresh distillate from rose petals.

SACCHARIN

1,2-Benzisothiazol-3(2H)-one, 1,1-dioxide; Gluside; *o*-Benzosulfimide



1,2-Benzisothiazolin-3-one 1,1-dioxide [81-07-2] $C_7H_5NO_3S$ (183.18).

Preparation—Toluene is reacted with chlorosulfonic acid to form *o*-toluenesulfonyl chloride, which is converted to the sulfonamide with ammonia. The methyl group then is oxidized with dichromate, yielding *o*-sulfamoylbenzoic acid, which, when heated, forms the cyclic imide.

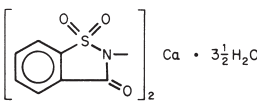
Description—White crystals or a white crystalline powder; odorless or with a faint aromatic odor; in dilute solution it is intensely sweet; solutions are acid to litmus; melts between 226 and 230°C.

Solubility—1 g in 290 mL water, 31 mL alcohol, or 25 mL boiling water; slightly soluble in chloroform or ether; readily dissolved by dilute solution of ammonia, solutions of alkali hydroxides, or solutions of alkali carbonates, with the evolution of CO_2.

Uses—A sweetening agent in Aromatic Cascara Sagrada Fluidextract and highly alcoholic preparations. It is an intensely sweet substance. A 60-mg portion is equivalent in sweetening power to approximately 30 g of sucrose. It is used as a sweetening agent in vehicles, canned foods, and beverages and in diets for diabetics to replace the sucrose. The relative sweetening power of saccharin is increased by dilution.

SACCHARIN CALCIUM

1,2-Benzisothiazol-3(2H)-one, 1,1-dioxide, calcium salt, hydrate (2:7) Calcium *o*-Benzosulfimide



1,2-Benzisothiazolin-3-one 1,1-dioxide calcium salt hydrate (2:7) [6381-91-5] $C_{14}H_8CaN_2O_6S_2 \bullet 31/2\ H_2O$ (467.48); anhydrous [6485-34-3] (404.43).

Preparation—Saccharin is reacted with a semimolar quantity of calcium hydroxide in aqueous medium, and the resulting solution is concentrated to crystallization.

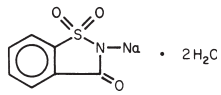
Description—White crystals or a white, crystalline powder; odorless or with a faint aromatic odor; and an intensely sweet taste even in dilute solutions; in dilute solution it is about 300 times as sweet as sucrose.

Solubility—1 g in 2.6 mL water or 4.7 mL alcohol.

Uses See Saccharin.

SACCHARIN SODIUM

1,2-Benzisothiazol-3(2H)-one, 1,1-dioxide, sodium salt, dihydrate; Soluble Saccharin; Soluble Gluside; Sodium *o*-Benzosulfimide



1,2-Benzisothiazolin-3-one 1,1-dioxide sodium salt dihydrate [6155- 57-3] $C_7H_4NNaO_3S \bullet 2\ H_2O$ (241.19); *anhydrous* [128-44-9] (205.16).

Preparation—Saccharin is dissolved in an equimolar quantity of aqueous sodium hydroxide, and the solution is concentrated to crystallization.

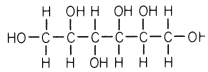
Description—White crystals or a white crystalline powder; odorless or with a faint aromatic odor, and an intensely sweet taste even in dilute solutions; in dilute solution it is about 300 times as sweet as sucrose; in powdered form it usually contains about 1/3 the theoretical amount of water of hydration because of efflorescence.

Solubility—1 g in 1.5 mL water or 50 mL alcohol.

Uses—Same as Saccharin but has the advantage of being more soluble in neutral aqueous solutions.

SORBITOL

Sionin; Sorbit; D-Sorbitol; D-Glucitol; Sorbo



D-Glucitol [50-70-4] $C_6H_{14}O_6$ (182.17); it may contain small amounts of other polyhydric alcohols.

Preparation—Commercially by reduction (hydrogenation) of certain sugars, such as glucose.

Description—White, hygroscopic powder, granules, or flakes, with a sweet taste; the usual form melts about 96°C.

Solubility—1 g in about 0.45 mL of water; slightly soluble in alcohol, methanol, or acetic acid.

Uses—An osmotic diuretic given intravenously in 50% (w/v) solution to diminish edema, lower cerebrospinal pressure, or reduce intraocular pressure in glaucoma. It also is used as a laxative, sweetener, humectant, plasticizer and, in 70% (w/w) solution, as a vehicle.

Sorbitol Solution—a water solution containing, in each 100 g, 69 to 71 g of total solids consisting essentially of D-sorbitol and a small amount of mannitol and other isomeric polyhydric alcohols. The content of D-sorbitol $C_6H_8(OH)_6$ in each 100 g is not less than 64 g. Description: Clear, colorless, syrupy liquid, with a sweet taste and no characteristic odor; neutral to litmus; specific gravity not less than 1.285; refractive index at 20 1.455 to 1.465. Uses: It is not to be injected. It has been used as a replacement for propylene glycol and glycerin.

SPEARMINT

Spearmint Leaves; Spearmint Herb; Mint

The dried leaf and flowering top of *Mentha spicata* Linné (*Mentha viridis* Linné) (Common Spearmint) or of *Mentha cardiaca* Gerard ex Baker (Scotch Spearmint) (Fam *Labiatae*).

Fresh spearmint is used in preparing mint sauce and also the well-known mint julep. The volatile oil is the only constituent of importance in this plant; the yield is from 1/2 to 1%.

Uses A flavoring agent.

Spearmint Oil—the volatile oil distilled with steam from the fresh overground parts of the flowering plant *Mentha spicata* or *Mentha cardiaca;* contains not less than 55%, by volume, of $C_{10}H_{14}O$ (carvone, 150.22). The chief odoriferous constituent is the ketone l-carvone. American oil also contains dihydrocarveol acetate [$CH_3COOC_{10}H_{17}$], l-limonene [$C_{10}H_{16}$], a small amount of phellandrene [$C_{10}H_{16}$], and traces of esters of valeric and caproic acids.

Description and Solubility—Colorless, yellow, or greenish yellow liquid, with the characteristic odor and taste of spearmint; specific gravity 0.917 to 0.934. Soluble in 1 volume of 80% alcohol, but upon further dilution may become turbid. Uses: Primarily as a flavoring agent. It also has been used as a carminative in doses of 0.1 mL.

SUCROSE

α-D-Glucopyranoside, β-D-fructofuranosyl-, Sugar; Cane Sugar; Beet Sugar

Sucrose [57-50-1] $C_{12}H_{22}O_{11}$ (342.30); a sugar obtained from *Saccharum officinarum* Linné (Fam *Gramineae*), *Beta vulgaris* Linné (Fam *Chenopodiaceae*), and other sources. It contains no added substances.

Preparation—Commercially from sugar cane, beat root, and sorghum. Originally, sugar cane was the only source, but at present the root of *Beta vulgaris* is used largely in Europe, and to an increasing degree in this country, for making sucrose.

The sugar cane is crushed, and the juice amounting to about 80% is expressed with roller mills. The juice, after defecation with lime and removal of excess of lime by carbonic acid gas, is run into vacuum pans for concentration, and the saccharine juice is evaporated in this until it begins to crystallize. After the crystallization is complete, the warm mixture of crystals and syrup is run into centrifuges, in which the crystals of raw sugar are drained and dried. The syrup resulting as a by-product from raw sugar is known as molasses. Raw beet sugar is made by a similar process but is more troublesome to purify than that made from sugar cane.

The refined sugar from either raw cane or beet sugar is prepared by dissolving the raw sugar in water, clarifying, filtering, and finally decolorizing the solution by passing it through bone-black filters. The water-white solution finally is evaporated under reduced pressure to the crystallizing point and then forced to crystallize in small granules that are collected and drained in a centrifuge.

Description—Colorless or white crystals, crystalline masses or blocks, or a white, crystalline powder; odorless; sweet taste; stable in air; solutions neutral to litmus; melts with decomposition from 160 to 185°C; specific gravity of about 1.57; specific rotation at 20°C not less than +65.9; unlike the other official sugars (dextrose, fructose, and lactose), it does not reduce Fehling's solution even in hot solutions; also differs from these sugars in that it is darkened and charred by sulfuric acid in the cold, is fermentable, and in dilute aqueous solutions, it ferments into alcohol and eventually acetic acid.

Sucrose is hydrolyzed by dilute mineral acids, slowly in the cold and rapidly on heating, into one molecule each of dextrose or levulose. This process is known technically as inversion and the product is referred to as invert sugar the term inversion being derived from the change, through the hydrolysis, in the optical rotation from dextro of sucrose to levo of the hydrolyzed product. The enzyme invertase also hydrolyzes sucrose.

Solubility—1 g in 0.5 mL water, 170 mL alcohol, or slightly more than 0.2 mL boiling water; insoluble in chloroform or ether.

Uses—Principally as a pharmaceutical necessity for making syrups and lozenges. It gives viscosity and consistency to fluids.

Intravenous administration of hypertonic solutions has been employed chiefly to initiate osmotic diuresis. Such a procedure is not completely safe, and renal tubular damage may result, particularly in patients with existing renal pathology. Safer and more effective diuretics are available.

CONFECTIONER'S SUGAR

Sucrose ground together with corn starch to a fine powder; contains 95.0 to 97.0% sucrose.

Description—Fine, white, odorless powder; sweet taste; stable in air; specific rotation not less than +62.6.

Solubility—The sucrose portion is soluble in cold water; this is entirely soluble in boiling water.

Uses—A pharmaceutical aid as a tableting excipient and sweetening agent. See also Sucrose.

SYRUP—page 1071.

TOLU BALSAM

Tolu

A balsam obtained from *Myroxylon balsamum* (Linné) Harms (Fam *Leguminosae*).

Constituents—Up to 80% resin, about 7% volatile oil, 12 to 15% free cinnamic acid, 2 to 8% benzoic acid, and 0.05% vanillin. The volatile oil is composed chiefly of benzyl benzoate, and benzyl cinnamate, ethyl benzoate, ethyl cinnamate, a terpene called tolene (possibly identical with phellandrene), and the sesquiterpene alcohol farnesol also have been reported to be present.

Description—Brown or yellowish brown, plastic solid; transparent in thin layers and brittle when old, dried, or exposed to cold temperatures; pleasant, aromatic odor resembling that of vanilla and a mild, aromatic taste.

Solubility—Nearly insoluble in water or solvent hexane; soluble in alcohol, chloroform, or ether, sometimes with slight residue or turbidity.

Uses—A vehicle, flavoring agent, and stimulating expectorant as a syrup. It is also an ingredient of Compound Benzoin Tincture (page 1280).

Tolu Balsam Syrup [Syrup of Tolu; Tolu Syrup]—Preparation: Add tolu balsam tincture (50 mL, all at once) to magnesium carbonate (10 g) and sucrose (60 g) in a mortar, and mix intimately. Gradually add purified water (430 mL) with trituration, and filter. Dissolve the remainder of the sucrose (760 g) in the clear filtrate with gentle heating, strain the syrup while warm, and add purified water (qs) through the strainer to make the product measure 1000 mL. Mix thoroughly. Note: May be made also in the following manner: Place the remaining sucrose (760 g) in a suitable percolator, the neck of which nearly is filled with loosely packed cotton, moistened after packing with a few drops of water. Pour the filtrate, obtained as directed in the formula above, upon the sucrose, and regulate the outflow to a steady drip of percolate. When all of the liquid has run through, return portions of the percolate, if necessary, to dissolve all of the sucrose. Then pass enough purified water through the cotton to make the product measure 1000 mL. Mix thoroughly. Alcohol Content: 3 to 5%. Uses: Chiefly for its agreeable flavor in cough syrups. Dose: 10 mL.

Tolu Balsam Tincture [Tolu Tincture]—Preparation: With tolu balsam (200 g), prepare a tincture by Process M using alcohol as the menstruum. Alcohol Content: 77 to 83%. Uses: A balsamic preparation employed as an addition to expectorant mixtures; also used in the preparation of Tolu Balsam Syrup.

VANILLA

Vanilla Bean

The cured, full-grown, unripe fruit of *Vanilla planifolia* Andrews, often known in commerce as Mexican or Bourbon Vanilla, or of *Vanilla tahitensis* JW Moore, known in commerce as Tahiti Vanilla (Fam *Orchidaceae*); yields not less than 12% of anhydrous extractive soluble in diluted alcohol.

Constituents—Contains a trace of a volatile oil, fixed oil, 4% resin, sugar, vanillic acid, and about 2.5% vanillin (see below). This highest grade of vanilla comes from Madagascar; considerable quantities of the drug also are produced in Mexico.

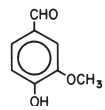
Uses—A flavor.

Note—Do not use if it has become brittle.

Vanilla Tincture [Extract of Vanilla]—Preparation: Add water (200 mL) to comminuted vanilla (cut into small pieces, 100 g) in a suitable covered container, and macerate during 12 hr, preferably in a warm place. Add alcohol (200 mL) to the mixture of vanilla and water, mix well, and macerate about 3 days. Transfer the mixture to a percolator containing sucrose (in coarse granules, 200 g), and drain; then pack the drug firmly, and percolate slowly, using diluted alcohol (qs) as the solvent. If the percolator is packed with an evenly distributed mixture of the comminuted vanilla, sucrose, and clean, dry sand, the increased surface area permits more efficient percolation. This tincture is unusual in that it is the only official one in which sucrose is specified as an ingredient. Alcohol Content: 38 to 42%. Uses: A flavoring agent. See Flavors, page 1061.

VANILLIN

Benzaldehyde, 4-hydroxy-3-methoxy-,



4-Hydroxy-3-methoxybenzaldehyde [121-33-5] $C_8H_8O_3$ (152.15).

Preparation—From vanilla, which contains 2 to 3%. It also is found in many other substances, including tissues of certain plants, crude beet sugar, asparagus, and even asafetida. Commercially, it is made synthetically. While chemically identical with the product obtained from the vanilla bean, flavoring preparations made from it never equal in flavor the preparation in which vanilla alone is used, because vanilla contains other odorous products. It is synthesized by oxidation processes from either coniferin or eugenol, by treating guaiacol with chloroform in the presence of an alkali, and by other methods.

Description—Fine, white to slightly yellow crystals, usually needle-like, with an odor and taste suggestive of vanilla; affected by light; solutions are acid to litmus; melts 81 to 83°C.

Solubility—1 g in about 100 mL water, about 20 mL glycerin, or 20 mL water at 80°C; freely soluble in alcohol, chloroform, ether, or solutions of the fixed alkali hydroxides.

Incompatibilities—Combines with glycerin, forming a compound that is almost insoluble in alcohol. It is decomposed by alkali and is oxidized slowly by the air.

Uses—Only as a flavor. Solutions of it sometimes are sold as a synthetic substitute for vanilla for flavoring foods, but it is inferior in flavor to the real vanilla extract.

WATER—page 1070.
WATER PURIFIED—page 1070.
WILD CHERRY SYRUP—page 1069.

OTHER FLAVORING AGENTS

Anise NF [Anise Seed; European Aniseed; Sweet Cumin]—The dried ripe fruit of *Pimpinella anisum* Linné. It contains about 1.75% of volatile oil. Uses: A flavor and carminative.

Ceylon Cinnamon—The dried inner bark of the shoots of coppiced trees of *Cinnamomum zeylanicum* Nees (Fam *Lauraceae*); contains, in each 100 g, not less than 0.5 mL volatile oil. Uses: A carminative and flavor.

Clove—The dried flower-bud of *Eugenia caryophyllus* (Sprengel) Bullock et Harrison (Fam *Myrtaceae*). It contains, in each 100 g, not less than 16 mL of clove oil. Uses: An aromatic in doses of 0.25 g and as a condiment in foods.

Coriander—The dried ripe fruit of *Coriandrum sativum* Linné (Fam *Umbelliferae*); yields not less than 0.25 mL volatile coriander oil/100 g. Uses: Seldom used alone, but sometimes is combined with other agents, chiefly as a flavor. It also is used as a condiment and flavor in cooking.

Eucalyptol [Cineol; Cajeputol] C10H18O (154.25)—Obtained from eucalyptus oil and from other sources. Colorless liquid, with a characteristic aromatic, distinctly camphoraceous odor and a pungent, cooling, spicy taste; 1 volume is soluble in 5 volumes of 60% alcohol; miscible with alcohol, chloroform, ether, glacial acetic acid, or fixed or volatile oils; insoluble in water. Uses: Primarily as a flavoring agent. Locally it is employed for its antiseptic effect in inflammations of the nose and throat and in certain skin diseases. It sometimes is used by inhalation in bronchitis.

Fennel [Fennel Seed]—The dried ripe fruit of cultivated varieties of *Foeniculum vulgare* Miller (Fam *Umbelliferae*); contains 4 to 6% of an oxygenated volatile oil and 10% of a fixed oil. Uses: A flavor and carminative.

Ginger NF [Zingiber]—The dried rhizome of *Zingiber officinale* Roscoe (Fam *Zingiberaceae*), known in commerce as Jamaica Ginger, African Ginger and Cochin Ginger. The outer cortical layers often are removed either partially or completely. Constituents: A pungent substance, gingerol; volatile oil (Jamaica Ginger, about 1%; African Ginger, 2 to 3%), containing the terpenes d-camphene and β-phellandrene and the sesquiterpene zingiberene; citral cineol and borneol. Uses: A flavoring agent. It formerly was employed in a dose of 600 mg as an intestinal stimulant and carminative in colic and in diarrhea.

Ginger Oleoresin—Yields 18 to 35 mL of volatile ginger oil/100 g of oleoresin. Preparation: Extract the oleoresin from ginger, in moderately coarse powder, by percolation, using either acetone, alcohol, or ether as the menstruum.

Glycyrrhiza Extract [Licorice Root Extract; Licorice]—An extract prepared from the rhizome and roots of species of *Glycyrrhiza* Tournefort ex Linné (Fam *Leguminosae*). Description: Brown powder or in flattened, cylindrical rolls, or in masses; the rolls or masses have a glossy black color externally and a brittle, sharp, smooth, conchoidal fracture; the extract has a characteristic sweet taste that is not more than very slightly acrid. Uses: A flavoring agent.

Lavender [Lavendula]—The flowers of *Lavandula spica* (*Lavandula officinalis* or *Lavandula vera*); contains a volatile oil with the principal constituent l-linalyl acetate. Uses: A perfume.

Lemon Peel USP, BP [Fresh Lemon Peel]—The outer yellow rind of the fresh ripe fruit of *Citrus limon* (Linné) Burmann filius (Fam *Rutaceae*); contains a volatile oil and hesperidin. Uses: A flavor.

Lemon Tincture USP [Lemon Peel Tincture]—Preparation: From lemon peel, which is the outer yellow rind of the fresh, ripe fruit of Citrus limon (Linné) Burmann filius (Fam Rutaceae), by Process M, 500 g of the peel being macerated in 900 mL alcohol, and the preparation being completed with alcohol to make the product measure 1000 mL. Use talc as the filtering medium. The white portion of the rind must not be used, as the proportion of oil, which is found only in the yellow rind, is reduced, and the bitter principle, hesperidin, introduced. Alcohol Content: 62 to 72%. Uses: A flavor, its fineness of flavor being ensured as it comes from the fresh fruit, and being an alcoholic solution it is more stable than the oil.

Myrcia Oil [Bay Oil; Oil of Bay]—The volatile oil distilled from leaves of *Pimenta racemosa* (Miller) JW Moore (Fam *Myrtaceae*); contains the phenolic compounds eugenol and chavicol. Uses: In the preparation of bay rum as a perfume.

Orange Oil, Bitter—The volatile oil obtained by expression from the fresh peel of the fruit of *Citrus aurantium* Linné (Fam *Rutaceae*); contains primarily *d*-limonene. Pale yellow liquid with a characteristic aromatic odor of the Seville orange; if it has a terebinthinate odor, it should not be dispensed; refractive index 1.4725 to 1.4755 at 20°C. It differs little from Orange Oil except for the botanical source. Miscible with anhydrous alcohol and with about 4 volumes alcohol. Uses: A flavor.

Orange Peel, Bitter [Bitter Orange; Curacao Orange Peel; Bigarade Orange]. The dried rind of the unripe but fully grown fruit of *Citrus aurantium* Linné (Fam *Rutaceae*). Constituents: The inner part of the peel from the bitter orange contains a volatile oil and the glycoside hesperidin ($C_{28}H_{34}O_{15}$). This, upon hydrolysis in the presence of H_2SO_4, yields hesperetin ($C_{16}H_{14}O_6$), rhamnose ($C_6H_{12}O_5$), and D-glucose ($C_6H_{12}O_6$). Uses: A flavoring agent. It has been used as a bitter.

Orange Peel, Sweet USP—The fresh outer rind of the non-artificially-colored, ripe fruit of *Citrus sinensis* (Linné) Osbeck (Fam *Rutaceae*); the white inner portion of the rind is to be excluded. Contains a volatile oil but no hesperidin, since the glycoside occurs in the white portion of the rind. Uses: A flavor.

Orris [Orris Root; Iris; Florentine Orris]—The peeled and dried rhizome of *Iris germanica* Linné, including its variety *florentina* Dykes (*Iris florentina* Linné), or of *Iris pallida* Lamarck (Fam *Iridaceae*); contains about 0.1 to 0.2% of a volatile oil (orris butter), myristic acid and the ketone irone; irone provides the fragrant odor of orris. Uses: A perfume.

Pimenta Oil [Pimento Oil; Allspice Oil]—The volatile oil distilled from the fruit of *Pimenta officinalis* Lindley (Fam *Myrtaceae*). Uses: A carminative and stimulant and also as a condiment in foods.

Rosemary Oil—The volatile oil distilled with steam from the fresh flowering tops of *Rosmarinus officinalis* Linné (Fam *Labiatae*); yields not less than 1.5% of esters calculated as bornyl acetate ($C_{12}H_{20}O_2$) and not less than 8% of total borneol ($C_{10}H_{18}O$), free and as esters. Constituents: The amount of esters, calculated as bornyl acetate, and of total borneol, respectively, varies somewhat with its geographical source. Cineol is present to the extent of about 19 to 25%, depending on the source. The terpenes d- and l-α-pinene, dipentene, and camphene, and the ketone camphor also occur in this oil. Description and Solubility: Colorless or pale yellow liquid, with the characteristic odor of rosemary and a warm, camphoraceous taste; specific gravity 0.894 to 0.912. Soluble in 1 volume of 90% alcohol, by volume, but upon further dilution may become turbid. Uses: A flavor and perfume, chiefly, in rubefacient liniments such as Camphor and Soap Liniment.

Sassafras—The dried bark of the root of *Sassafras albidum* (Nuttall) Nees (Fam *Lauraceae*). Uses: Principally because of its high content of volatile oil that serves to disguise the taste of disagreeable substances. An infusion (sassafras tea) formerly was used extensively as a home remedy, particularly in the southern states.

Sassafras Oil—The volatile oil distilled with steam from Sassafras. Uses: A flavor by confectioners, particularly in hard candies. Either the oil or safrol is used as a preservative in mucilage and library paste, being far superior to methyl salicylate for this purpose. Since the oil is antiseptic, it sometimes is employed in conjunction with other agents for local application in diseases of the nose and throat; safrol also is used in this way.

Wild Cherry [Wild Black Cherry Bark]—The carefully dried stem bark of *Prunus serotina* Ehrhart (Fam *Rosaceae*), free of borke and preferably having been collected in autumn. Constituents: A glucoside of d-mandelonitrile ($C_6H_5 \cdot CHOH \cdot CN$) known as prunasin, the enzyme emulsin, tannin, a bitter principle, starch, resin, etc. In the BP and the English literature this drug has been termed Virginian Prune—a literal but incorrect translation of the older botanical name, *Prunus virginiana*. Uses: A flavoring agent, especially in cough preparations. It is an ingredient in Wild Cherry Syrup. As with bitter almond, contact with water, in the presence of emulsin, results in the production of benzaldehyde and HCN. All preparations of wild cherry should be made without heat, to avoid destruction of the enzyme that is responsible for the production of the free active principles.

Diluting Agents

Diluting agents (vehicles or carriers) are indifferent substances that are used as solvents for active medicinals. They are of primary importance for diluting and flavoring drugs that are intended for oral administration, but a few such agents are designed specifically for diluting parenteral injections. The latter group is considered separately.

The expert selection of diluting agents has been an important factor in popularizing the specialties of compounding pharmacists. Since a large selection of diluting agents is available in a choice of colors and flavors, prescribers have the opportunity to make their own prescriptions more acceptable to the patient. The best diluting agent is usually the best solvent for the drug. Water-soluble substances, for example, should be flavored and diluted with an aqueous agent, and alcohol-soluble drugs with an alcoholic vehicle. Thus, the diluting agents presented herein are divided into three groups on the basis of their physical properties: aqueous, hydroalcoholic, and alcoholic.

AQUEOUS DILUTING AGENTS

Aqueous diluting agents include aromatic waters, syrups, and mucilages. Aromatic waters are used as diluting agents for

water-soluble substances and salts but cannot mask the taste of very disagreeable drugs. Some of the more common flavored aqueous agents and the official forms of water are listed below.

ORANGE FLOWER WATER

Stronger Orange Flower Water; Triple Orange Flower Water

A saturated solution of the odoriferous principles of the flowers of *Citrus aurantium* Linné (Fam *Rutaceae*), prepared by distilling the fresh flowers with water and separating the excess volatile oil from the clear, water portion of the distillate.

Description—Should be nearly colorless, clear, or only faintly opalescent; the odor should be that of the orange blossoms; it must be free from empyreuma, mustiness, and fungoid growths.

Uses—A vehicle flavor and perfume in syrups, elixirs, and solutions.

PEPPERMINT WATER

A clear, saturated solution of peppermint oil in purified water, prepared by one of the processes described under Aromatic Waters (page 749).

Uses—A carminative and flavored vehicle.

TOLU BALSAM SYRUP—page 1068.

WATER

Water [7732-18-5] H_2O (18.02).

Drinking water, which is subject to EPA regulations with respect to drinking water and which is delivered by the municipal or other local public system or drawn from a private well or reservoir, is the starting material for all forms of water covered by Pharmacopeial monographs.

Drinking water may be used in the preparation of USP drug substances (eg, in the extraction of certain vegetable drugs and in the manufacture of a few preparations used externally) but not in the preparation of dosage forms or in the preparation of reagents or test solutions. It is no longer the subject of a separate monograph (in the USP), inasmuch as the cited standards vary from one community to another and generally are beyond the control of private parties or corporations.

PURIFIED WATER

Water obtained by distillation, ion-exchange treatment, reverse osmosis, or any other suitable process; contains no added substance.

Caution—Do not use this in preparations intended for parenteral administration. For such purposes, use Water for Injection, Bacteriostatic Water for Injection, or Sterile Water for Injection, page 1070.

Preparation—From water complying with EPA regulations with respect to drinking water. A former official process for water, when prepared by distillation, is given below. The pharmacist who is preparing sterile solutions and must have freshly distilled water of exceptionally high grade, not only free from all bacterial or other microscopic growths but also free from the products of metabolic processes resulting from the growth of such organisms in the water, advantageously may follow this plan. The metabolic products commonly are spoken of as pyrogens and usually consist of complex organic compounds that cause febrile reactions if present in the solvent for parenteral medicinal substances.

DISTILLATION PROCESS

Water	1000 vol.
To make	750 vol.

Distill the water from a suitable apparatus provided with a block-tin or glass condenser. Collect the first 100 volumes and reject this portion. Then collect 750 volumes and keep the distilled water in glass-stoppered bottles that have been rinsed with steam or very hot distilled water immediately before being filled. The first 100 volumes are discarded to eliminate foreign volatile substances found in ordinary water, and only 750 volumes are collected, since the residue in the still contains concentrated dissolved solids.

Description—Colorless, clear liquid, without odor or taste.

Uses—A pharmaceutical aid (vehicle and solvent). It must be used in compounding dosage forms for internal (oral) administration as well as sterile pharmaceuticals applied externally, such as collyria and dermatological preparations, but these must be sterilized before use.

Whenever water is called for in official tests and assays, this must be used.

WATER FOR INJECTION

Water purified by distillation or by reverse osmosis. It contains no added substance.

Caution—It is intended for use as a solvent for the preparation of parenteral solutions. For parenteral solutions that are prepared under aseptic conditions and are not sterilized by appropriate filtration or in the final container, first render it sterile and thereafter protect it from microbial contamination.

Description—Clear, colorless, odorless liquid.

Uses—Vehicle and solvent

BACTERIOSTATIC WATER FOR INJECTION

Sterile water for injection containing one or more suitable antimicrobial agents

Note—Use it with due regard for the compatibility of the antimicrobial agent or agents it contains with the particular medicinal substance that is to be dissolved or diluted.

Uses—Sterile vehicle for parenteral preparations.

STERILE WATER FOR INJECTION

WATER FOR PARENTERALS

Water for injection sterilized and suitably packaged. It contains no antimicrobial agent or other added substance.

Description—Clear, colorless, odorless liquid.

Uses—For the preparation of all aqueous parenteral solutions including those used in animal assays.

STERILE WATER FOR IRRIGATION

Water for injection that has been sterilized and suitably packaged. It contains no antimicrobial agent or other added substance.

Description—Clear, colorless, odorless liquid.

Uses—An irrigating solution.

SYRUPS USED AS DILUTING AGENTS

Syrups are useful as diluting agents for water-soluble drugs and act both as solvents and flavoring agents. The flavored syrups usually consist of simple syrup (85% sucrose in water) containing appropriate flavoring substances. Glycyrrhiza Syrup is an excellent vehicle for saline substances because of its colloidal properties, sweet flavor, and lingering taste of licorice. Acacia Syrup is valuable in disguising the taste of urea. Fruit syrups are especially effective for masking sour tastes. Aromatic Eriodictyon Syrup is the diluting agent of choice for masking the bitter taste of alkaloids. Cocoa Syrup and Cherry Syrup are good general flavoring agents.

ACACIA SYRUP

Acacia, granular or powdered	100 g
Sodium Benzoate	1 g
Vanilla Tincture	5 mL
Sucrose	800 g
Purified Water, a sufficient quantity to make	1000 mL

Mix the acacia, sodium benzoate, and sucrose; then add 425 mL of purified water and mix well. Heat the mixture on a steam bath until solution is completed. When cool, remove the scum, and add the vanilla tincture and sufficient purified water to make the product measure 1000 mL and strain, if necessary.

Uses—A flavored vehicle and demulcent.

CHERRY SYRUP

Syrupus Cerasi

Cherry Juice	475 mL
Sucrose	800 g
Alcohol	20 mL
Purified Water, a sufficient quantity to make	1000 mL

Dissolve the sucrose in cherry juice by heating on a steam bath, cool, and remove the foam and floating solids. Add the alcohol and sufficient purified water to make 1000 mL and mix.

Alcohol Content—1 to 2%.

Uses—A pleasantly flavored vehicle that is particularly useful in masking the taste of saline and sour drugs.

COCOA SYRUP

Cacao Syrup; Chocolate-flavored Syrup; Chocolate Syrup

Cocoa	180 g
Sucrose	600 g
Liquid Glucose	180 g
Glycerin	50 mL
Sodium Chloride	2 g
Vanillin	0.2 g
Sodium Benzoate	1 g
Purified Water, a sufficient quantity to make	1000 mL

Mix the sucrose and the cocoa, and to this mixture gradually add a solution of the liquid glucose, glycerin, sodium chloride, vanillin, and sodium benzoate in 325 mL of hot purified water. Bring the entire mix-

ture to a boil, and maintain at boiling temperature for 3 min. Allow to cool to room temperature, and add sufficient purified water to make the product measure 1000 mL.

Note—Cocoa containing not more than 12% nonvolatile, ether- soluble, extractive (fat) yields a syrup having a minimum tendency to separate. Breakfast cocoa contains over 22% fat.

Uses—A pleasantly flavored vehicle.

SYRUP

Simple Syrup

Sucrose	850 g
Purified Water, a sufficient quantity, to make	1000 mL

May be prepared by using boiling water or, preferably, without heat, by the following process:

Place the sucrose in a suitable percolator the neck of which is nearly filled with loosely packed cotton, moistened, after packing, with a few drops of water. Pour carefully about 450 mL of purified water upon the sucrose, and regulate the outflow to a steady drip of percolate. Return the percolate, if necessary, until all of the sucrose has dissolved. Then wash the inside of the percolator and the cotton with sufficient purified water to bring the volume of the percolate to 1000 mL, and mix.

Specific Gravity—Not less than 1.30.

Uses—A sweet vehicle, sweetening agent, and as the basis for many flavored and medicated syrups.

OTHER SYRUPS USED AS DILUTING AGENTS

Glycyrrhiza Syrup USP [Licorice Syrup]—Preparation: Add fennel oil (0.05 mL) and anise oil (0.5 mL) to glycyrrhiza fluidextract (250 mL) and agitate until mixed. Then add syrup (qs) to make the product measure 1000 mL, and mix. Alcohol Content: 5 to 6%. Incompatibilities: The characteristic flavor is destroyed by acids because of precipitation of the glycyrrhizin. Uses: A flavored vehicle, especially adapted to the administration of bitter or nauseous substances.

Hydriodic Acid Syrup—Contains, in each 100 mL, 1.3 to 1.5 g HI (127.91). Preparation: Mix diluted hydriodic acid (140 mL) with purified water (550 mL), and dissolve dextrose (450 g) in this mixture by agitation. Add purified water (qs) to make the product measure 1000 mL, and filter. Caution: It must not be dispensed if it contains free iodine, as evidenced by a red coloration. Description: Transparent, colorless, or not more than pale straw-colored, syrupy liquid; odorless, with a sweet, acidulous taste; specific gravity about 1.18; hydriodic acid is decomposed easily in simple aqueous solution (unless protected by hypophosphorous acid), free iodine being liberated, and if taken internally, when in this condition, it is irritating to the alimentary tract. The dextrose used in this syrup should be of the highest grade obtainable.

Incompatibilities—The reactions of the acids as well as those of the water-soluble iodide salts. Oxidizing agents liberate iodine; alkaloids may be precipitated. Uses: Traditionally as a vehicle for expectorant drugs. Its therapeutic properties are those of the iodides. Dose: Usual, 5 mL.

Wild Cherry Syrup USP—Preparation: Pack wild cherry (in coarse powder, 150 g), previously moistened with water (100 mL), in a cylindrical percolator, and add water (qs) to leave a layer of it above the powder. Macerate for 1 hr, then proceed with rapid percolation, using added water, until 400 mL of percolate is collected. Filter the percolate, if necessary, add sucrose (675 g) and dissolve it by agitation, then add glycerin (150 mL), alcohol (20 mL), and water (qs) to make the product measure 1000 mL. Strain if necessary. It may be made also in the following manner: The sucrose may be dissolved by placing it in a second percolator as directed for preparing Syrup, and allowing the percolate from the wild cherry to flow through it and into a graduated vessel containing the glycerin and alcohol, until the total volume measures 1000 mL. Note: Heat is avoided, lest the enzyme emulsin be inactivated. If this should happen, the preparation would contain no free HCN, upon which its action as a sedative for coughs mainly depends. For a discussion of the chemistry involved, see Wild Cherry. Alcohol Content: 1 to 2%. Uses: Chiefly as a flavored vehicle for cough syrups.

MUCILAGES USED AS DILUTING AGENTS

Mucilages are also suitable as diluting agents for water-soluble substances, and are especially useful in stabilizing suspensions and emulsions.

The following mucilage used for this purpose is described under Emulsifying and Suspending Agents.

ACACIA MUCILAGE—page 1072.

HYDROALCOHOLIC DILUTING AGENTS

Hydroalcoholic diluting agents are suitable for drugs soluble in either water or diluted alcohol. The most important agents in this group are the elixirs. These solutions contain approximately 25% alcohol. Medicated elixirs that have therapeutic activity in their own right are not included in this section. Listed below are the common, non-medicated elixirs that are used purely as diluting agents or solvents for drugs.

AROMATIC ELIXIR

Simple Elixir

Orange Oil	2.4 mL
Lemon Oil	0.6 mL
Coriander Oil	0.24 mL
Anise Oil	0.06 mL
Syrup	375 mL
Talc	30 g
Alcohol, Purified Water, each, a sufficient quantity, to make	1000 mL

Dissolve the oils in alcohol to make 250 mL. To this solution add the syrup in several portions, agitating vigorously after each addition, and afterward add, in the same manner, the required quantity of purified water. Mix the talc with the liquid, and filter through a filter wetted with diluted alcohol, returning the filtrate until a clear liquid is obtained.

Alcohol Content—21 to 23%.

Uses—A pleasantly flavored vehicle, employed in the preparation of many other elixirs. The chief objection to its extensive use is the high alcohol content (about 22%), which at times may counteract the effect of other medicines.

OTHER HYDROALCOHOLIC DILUTING AGENTS

Glycyrrhiza Elixir [Elixir Adjuvants; Licorice Elixir]—Preparation: Mix glycyrrhiza fluidextract (125 mL) and aromatic elixir (875 mL) and filter. Alcohol Content: 21 to 23%. Uses: A flavored vehicle.

FLAVORED ALCOHOLIC SOLUTIONS

Flavored alcoholic solutions of high alcoholic concentration are useful as flavors to be added in small quantities to syrups or elixirs. The alcohol content of these solutions is approximately 50%. There are two types of flavored alcoholic solutions: tinctures and spirits. Only non-medicated tinctures and spirits are used as flavoring agents.

LEMON TINCTURE—page 1069.
ORANGE SPIRIT, COMPOUND—page 1066.
ORANGE PEEL, SWEET, TINCTURE—page 1066.

DILUTING AGENTS FOR INJECTIONS

Injections are liquid preparations, usually solutions or suspensions of drugs, intended to be injected through the skin into the body. Diluting agents used for these preparations may be aqueous or non-aqueous and must meet the requirements for sterility and also of the pyrogen test. Aqueous diluting agents include such preparations as Sterile Water for Injection and various sterile, aqueous solutions of electrolytes and/or dextrose. Non-aqueous diluting agents are generally fatty oils of vegetable origin, fatty esters, and polyols such as propylene glycol and polyethylene glycol. These agents are used to dissolve or dilute oil-soluble substances and to suspend water-soluble substances when it is desired to decrease the rate of absorption and, hence, prolong the duration of action of the drug substances. Preparations of this type are given intramuscularly. See *Parenteral Preparations,* page 802.

CORN OIL

Maize Oil

The refined fixed oil obtained from the embryo of *Zea mays* Linné (Fam *Gramineae*).

Preparation—Expressed from the Indian corn embryos or germs separated from the grain in starch manufacture.

Description—Clear, light yellow, oily liquid with a faint characteristic odor and taste; specific gravity 0.914 to 0.921.

Solubility—Slightly soluble in alcohol; miscible with ether, chloroform, benzene, or solvent hexane.

Uses—Main official use is as a solvent and vehicle for injections. It is used as an edible oil substitute for solid fats in the management of hypercholesterolemia. Other uses include making soaps and for burning.

It is a semidrying oil and therefore unsuitable for lubricating or mixing paint.

COTTONSEED OIL

Cotton Seed Oil; Cotton Oil

The refined fixed oil obtained from the seed of cultivated plants of various varieties of *Gossypium hirsutum* Linné or of other species of *Gossypium* (Fam *Malvaceae*).

Preparation—Cotton seeds contain about 15% oil. The testae of the seeds are first separated, and the kernels are subjected to high pressure in hydraulic presses. The crude oil thus has a bright red to blackish red color. It requires purification before it is suitable for medicinal or food purposes.

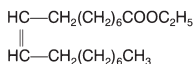
Description—Pale yellow, oily liquid with a bland taste; odorless or nearly so; particles of solid fat may separate below 10°C; solidifies at about 0 to −5°C; specific gravity 0.915 to 0.921.

Solubility—Slightly soluble in alcohol; miscible with ether, chloroform, solvent hexane, or carbon disulfide.

Uses—Officially as a solvent and vehicle for injections. It is sometimes taken orally as a mild cathartic in a dose of 30 mL or more. Taken internally, digestible oils retard gastric secretion and motility and increase the caloric intake. It also is used in the manufacture of soaps, oleomargarine, lard substitutes, glycerin, lubricants, and cosmetics.

ETHYL OLEATE

***(Z)*−9−Octadecenoic acid, ethyl ester**



Ethyl oleate [111−62−6] $C_{20}H_{38}O_2$ (310.52).

Preparation—Among other ways, by reacting ethanol with oleoyl chloride in the presence of a suitable dehydrochlorinating agent.

Description—Mobile, practically colorless liquid, with an agreeable taste; specific gravity 0.866 to 0.874; acid value not greater than 0.5; iodine value 75 to 85; sterilized by heating at 150°C for 1 hr; properties similar to those of almond and arachis oils, but is less viscous and more rapidly absorbed by the tissues; boils about 207°C.

Solubility—Does not dissolve in water; miscible with vegetable oils, mineral oil, alcohol, or most organic solvents.

Uses—A vehicle for certain intramuscular injectable preparations.

PEANUT OIL

Arachis Oil; Groundnut Oil; Nut Oil; Earth−Nut Oil

The refined fixed oil obtained from the seed kernels of one or more of the cultivated varieties of *Arachis hypogaea* Linné (Fam *Leguminosae*).

Description—Colorless or pale yellow, oily liquid, with a characteristic nutty odor and a bland taste; specific gravity 0.912 to 0.920.

Solubility—Very slightly soluble in alcohol; miscible with ether, chloroform, or carbon disulfide.

Uses—A solvent in preparing oil solutions for injection. It also is used for making liniments, ointments, plasters, and soaps, as a substitute for olive oil.

SESAME OIL

Teel Oil; Benne Oil; Gingili Oil

The refined fixed oil obtained from the seed of one or more cultivated varieties of *Sesamum indicum* Linné (Fam *Pedaliaceae*).

Description—Pale yellow, almost odorless, oily liquid with a bland taste; specific gravity 0.916 to 0.921.

Solubility—Slightly soluble in alcohol; miscible with ether, chloroform, solvent hexane, or carbon disulfide.

Uses—A solvent and vehicle in official injections. It is used much like olive oil both medicinally and for food. It does not readily turn rancid. It also is used in the manufacture of cosmetics, iodized oil, liniments, ointments, and oleomargarine.

EMULSIFYING AND SUSPENDING AGENTS

An emulsion is a two−phase system in which one liquid is dispersed in the form of small globules throughout another liquid that is immiscible with the first liquid. Emulsions are formed and stabilized with the help of emulsifying agents, which are surfactants and/or viscosity−producing agents. A suspension is defined as a preparation containing finely divided insoluble material suspended in a liquid medium. The presence of a suspending agent is required to overcome agglomeration of the dispersed particles and to increase the viscosity of the medium so that the particles settle more slowly. Emulsifying and suspending agents are used extensively in the formulation of elegant pharmaceutical preparations for oral, parenteral, and external use. For the theoretical and practical aspects of emulsions the interested reader is referred to Chapter 39 *(Solutions, Emulsions, Suspensions, and Extracts)*.

ACACIA

Gum Arabic

The dried gummy exudate from the stems and branches of *Acacia senegal* (Linné) Willdenow or of other related African species of *Acacia* (Fam *Leguminosae*).

Constituents—Principally calcium, magnesium, and potassium salts of the polysaccharide *arabic acid,* which on acid hydrolysis yields L-arabinose, L-rhamnose, D-galactose, and an aldobionic acid containing D-glucuronic acid and D-galactose.

Description—Acacia: Spheroidal tears up to 32 mm in diameter or angular fragments of white to yellowish white color; translucent or somewhat opaque; very brittle; almost odorless; produces a mucilaginous sensation on the tongue. Flake Acacia: White to yellowish white, thin flakes. Powdered Acacia: White to yellowish white, angular microscopic fragments. Granular Acacia: White to pale yellowish white, fine granules. Spray-dried Acacia: White to off-white compacted microscopic fragments or whole spheres.

Solubility—Insoluble in alcohol, but almost completely soluble in twice its weight of water at room temperature; the resulting solution flows readily and is acid to litmus.

Incompatibilities—Alcohol or alcoholic solutions precipitate acacia as a stringy mass when the alcohol amounts to more than about 35% of the total volume. Solution is effected by dilution with water. The mucilage is destroyed through precipitation of the acacia by heavy metals. Borax also causes a precipitation that is prevented by glycerin. It contains calcium and, therefore, possesses the incompatibilities of this ion.

It contains a peroxidase that acts as an oxidizing agent and produces colored derivatives of aminopyrine, antipyrine, cresol, guaiacol, phenol, tannin, thymol, vanillin, and other substances. Among the alkaloids affected are atropine, apomorphine, cocaine, homatropine, hyoscyamine, morphine, physostigmine, and scopolamine. A partial destruction of the alkaloid occurs in the reaction. Heating the solution of acacia for a few minutes at 100°C destroys the peroxidase and the color reactions are avoided.

Uses—Extensively as a suspending agent for insoluble substances in water, in the preparation of emulsions and for making pills and troches.

It is used for its demulcent action in inflammations of the throat or stomach.

Its solutions should not be used as a substitute for serum protein in the treatment of shock and as a diuretic in hypoproteinemic edema, since it produces serious syndromes that may result in death.

Acacia Mucilage [Mucilage of Gum Arabic]—Preparation: Place acacia (in small fragments, 350 g) in a graduated bottle having a wide mouth and a capacity not greatly exceeding 1000 mL, wash the drug with cold purified water, allow it to drain, and add enough warm purified water in which benzoic acid (2 g) has been dissolved, to make the product measure 1000 mL. After stoppering, lay the bottle on its side, rotate it occasionally, and when the acacia has dissolved, strain the mucilage. It also may be prepared as follows: dissolve benzoic acid (2 g) in purified water (400 mL) with the aid of heat, and add the solution to powdered or granular acacia (350 g), in a mortar, triturating until the acacia is dissolved. Then add sufficient purified water to make the product measure 1000 mL, and strain if necessary. This second method is primarily for extemporaneous preparation. Uses: A demulcent and a suspending agent. It also has been employed as an excipient in making pills and troches and as an emulsifying agent for cod liver oil and other

substances. Caution—It must be free from mold or any other indication of decomposition.

AGAR

Agar-Agar; Vegetable Gelatin; Gelosa; Chinese or Japanese Gelatin

The dried, hydrophilic, colloidal substance extracted from *Gelidium cartilagineum* (Linné) Gaillon (Fam *Gelidiaceae), Gracilaria confervoides* (Linné) Greville (Fam *Sphaerococcaceae*) and related red algae (Class *Rhodophyceae*).

Constituents—Chiefly of the calcium salt of a galactan mono-(acid sulfate).

Description—Usually in bundles of thin, membranous, agglutinated strips or in cut, flaked, or granulated forms; may be weak yellowish orange, yellowish gray to pale yellow or colorless; tough when damp, brittle when dry; odorless or with a slight odor; produces a mucilaginous sensation on the tongue. Also supplied as a white to yellowish white or pale-yellow powder.

Solubility—Insoluble in cold water; soluble in boiling water.

Incompatibilities—Like other gums, it is dehydrated and precipitated from solution by alcohol. Tannic acid causes precipitation; electrolytes cause partial dehydration and decrease in viscosity of sols.

Uses—A relatively ineffective bulk-producing laxative used in a variety of proprietary cathartics. In mineral oil emulsions it acts as a stabilizer. It also is used in culture media for bacteriological work and in the manufacture of ice cream, confectioneries, etc.

ALGINIC ACID

Alginic acid [9005-32-7] (average equivalent weight 200); a hydrophilic colloidal carbohydrate extracted with dilute alkali from various species of brown seaweeds *(Phaeophyceae)*.

Preparation—Precipitates when an aqueous solution of *Sodium Alginate* is treated with mineral acid.

Description—White to yellowish white, fibrous powder; odorless or practically odorless, and tasteless; pH (3 in 100 dispersion in water) 1.5 to 3.5; pK_a (0.1 N NaCl, 20°C) 3.42.

Solubility—Insoluble in water or organic solvents; soluble in alkaline solutions.

Uses—A tablet binder and emulsifying agent. It is used as a sizing agent in the paper and textile industries.

SODIUM ALGINATE

Alginic acid, sodium salt; Algin; Kelgin; Manucol; Norgine

Sodium alginate [9005-38-3] (average equivalent weight 220); the purified carbohydrate product extracted from brown seaweeds by the use of dilute alkali. It consists chiefly of the sodium salt of alginic acid, a polyuronic acid composed of beta-D-mannuronic acid residues linked so that the carboxyl group of each unit is free while a glycosidic linkage shields the aldehyde group.

Description—Nearly odorless and tasteless, coarse or fine powder, yellowish white in color.

Solubility—Dissolves in water, forming a viscous, colloidal solution; insoluble in alcohol or in hydroalcoholic solutions in which the alcohol content is greater than about 30% by weight; insoluble in chloroform, ether, or acids, when the pH of the solution becomes lower than about 3.

Uses—A thickening and emulsifying agent. This property makes it useful in a variety of areas. For example, it is used to impart smoothness and body to ice cream and to prevent formation of ice particles.

BENTONITE

Wilhinite; Soap Clay; Mineral Soap

Bentonite [1302-78-9]; a native, colloidal, hydrated aluminum silicate.

Occurrence—Bentonite is found in midwestern United States and Canada. Originally called Taylorite after its discoverer in Wyoming, its name was changed to bentonite after its discovery in the Fort Benton formation of the Upper Cretaceous of Wyoming.

Description—Very fine, odorless powder with a slightly earthy taste, free from grit; the powder is nearly white, but may be pale buff or cream colored.

The US Geological Survey has defined bentonite as transported stratified clay formed by the alteration of volcanic ash shortly after deposition. Chemically, it is $Al_2O_3 \cdot 4SiO_2 \cdot H_2O$ plus other minerals as impurities. It consists of colloidal crystalline plates, of less than microscopic dimensions in thickness, and of colloidal dimensions in breadth. This fact accounts for the extreme swelling that occurs when it is placed in water, since the water penetrates between an infinite number of plates. A good specimen swells 12 to 14 times its volume.

Solubility—Insoluble in water or acids, but it has the property of adsorbing large quantities of water, swelling to approximately 12 times its original volume, and forming highly viscous thixotropic suspensions or gels. This property makes it highly useful in pharmacy. Its gel-forming property is augmented by the addition of small amounts of alkaline substances, such as magnesium oxide. It does not swell in organic solvents.

Incompatibilities—Acids and acid salts decrease its water- absorbing power and thus cause a breakdown of the magma. Suspensions are most stable at a pH above 7.

Uses—A protective colloid for the stabilization of suspensions. It also has been used as an emulsifier for oil and as a base for plasters, ointments, and similar preparations.

Bentonite Magma—Preparation: Sprinkle bentonite (50 g), in portions, on hot purified water (800 g), allowing each portion to become thoroughly wetted without stirring. Allow it to stand with occasional stirring for 24 hr. Stir until a uniform magma is obtained, add purified water to make 1000 g, and mix. The magma may be prepared also by mechanical means such as by use of a blender, as follows: Place purified water (about 500 g) in the blender, and while the machine is running, add bentonite (50 g). Add purified water to make up to about 1000 g or up to the operating capacity of the blender. Blend the mixture for 5 to 10 min, add purified water to make 1000 g, and mix. Uses: A suspending agent for insoluble medicaments.

CARBOMER

Carboxy polymethylene; polyacrylic acid; acrylic acid polymer; carboxyvinyl polymer

A synthetic high-molecular-weight cross-linked polymer of acrylic acid; contains 56% to 68% of carboxylic acid (-COOH) groups. The viscosity of a neutralized preparation (2.5 g/500 mL water) is 30,000 to 40,000 centipoise.

Description—White, fluffy powder with a slight, characteristic odor; hygroscopic; pH (1 in 100 dispersion) about 3; specific gravity about 1.41.

Solubility—neutralized with alkali hydroxides or amines); dissolves in water, alcohol, or glycerin.

Uses—A thickening, suspending, dispersing and emulsifying agent for pharmaceuticals, cosmetics, waxes, paints, and other industrial products.

CARRAGEENAN

Carrageenan [9000-07-1]; Chondrus; Irish Moss

Preparation—The hydrocolloid extracted with water or aqueous alkali from certain red seaweeds of the class *Rhodophyceae,* and separated from the solution by precipitation with alcohol (methanol, ethanol, or isopropanol) or by drum-roll drying or freezing.

Constituents—It is a variable mixture of potassium, sodium, calcium, magnesium, and ammonium sulfate esters of galactose and 3,6-anhydrogalactose copolymers, the hexoses being alternately linked α-1,3 and β-1,4 in the polymer. The three main types of copolymers present are *kappa*-carrageenan, *iota*-carrageenan, and *lambda*-carrageenan, which differ in the composition and manner of linkage of monomeric units and the degree of sulfation (the ester sulfate content for carrageenans varies from 18% to 40%). *Kappa*-carrageenan and *iota*-carrageenan are the gelling fractions; *lambda*-carrageenan is the nongelling fraction. The gelling fractions may be separated from the nongelling fraction by addition of potassium chloride to an aqueous solution of carrageenan. Carrageenan separated by drum-roll drying may contain mono- and di-glycerides or up to 5% of polysorbate 80, used as roll-stripping agents.

Description—Yellow-brown to white, coarse to fine powder; odorless; tasteless, producing a mucilaginous sensation on the tongue.

Solubility—All carrageenans hydrate rapidly in cold water, but only *lambda*-carrageenan and sodium carrageenans dissolve completely. Gelling carrageenans require heating to about 80°C for complete solution when potassium and calcium ions are present.

Uses—In the pharmaceutical and food industries as an emulsifying, suspending, and gelling agent.

CARBOXYMETHYLCELLULOSE SODIUM

Carbose D; Carboxymethocel S; CMC; Cellulose Gum

Cellulose, carboxymethyl ether, sodium salt [9004-32-4]; contains 6.5 to 9.5% of sodium (Na), calculated on the dried basis. It is available in several viscosity types: low, medium, high, and extra high.

Description—White to cream-colored powder or granules; the powder is hygroscopic; pH (1 in 100 aqueous solution) about 7.5.

Solubility—Easily dispersed in water to form colloidal solutions; insoluble in alcohol, ether, or most other organic solvents.

Uses—Suspending agent, tablet excipient, or viscosity-increasing agent. In tablet

POWDERED CELLULOSE

Cellulose [9004-34-6] $(C_6H_{10}O_5)_n$; purified, mechanically disintegrated cellulose prepared by processing alpha cellulose obtained as a pulp from fibrous plant materials.

Description—White, odorless substance, consisting of fibrous particles, which may be compressed into self-binding tablets that disintegrate rapidly in water; exists in various grades, exhibiting degrees of fineness ranging from a free-flowing dense powder to a coarse, fluffy, non-flowing material; pH (supernatant liquid of a 10 g/90 mL aqueous suspension after 1 hr) 5 to 7.5.

Solubility—Insoluble in water, dilute acids, or nearly all organic solvents; slightly soluble in NaOH solution (1 in 20).

Uses—*T*ablet diluent, adsorbent, or suspending agent.

CETYL ALCOHOL—page 1078.

CHOLESTEROL

Cholest-5-en-3-ol, (3β)-, Cholesterin

Cholest-5-en-3β-ol [57-88-5] $C_{27}H_{46}O$ (386.66).

A steroid alcohol widely distributed in the animal organism. In addition to cholesterol and its esters, several closely related steroid alcohols occur in the yolk of eggs, the brain, milk, fish oils, wool fat (10 to 20%), etc. These closely resemble it in properties. One of the methods of commercial production involves extraction of it from the unsaponifiable matter in the spinal cord of cattle, using petroleum benzin. Wool fat also is used as a source.

Description—White or faintly yellow, almost odorless pearly leaflets or granules; usually acquires a yellow to pale tan color on prolonged exposure to light or to elevated temperatures; melts 147 to 150°C.

Solubility—Insoluble in water; 1 g slowly dissolves in 100 mL alcohol or about 50 mL dehydrated alcohol; soluble in acetone, hot alcohol, chloroform, dioxane, ether, ethyl acetate, solvent hexane, or vegetable oils.

Uses—To enhance incorporation and emulsification of medicinal products in oils or fats. It is a pharmaceutical necessity for Hydrophilic Petrolatum in which it enhances water-absorbing capacity. See Chapter 21.

DOCUSATE SODIUM—page 1308.

GELATIN

White Gelatin

A product obtained by the partial hydrolysis of collagen derived from the skin, white connective tissues, and bones of animals. Gelatin derived from an acid-treated precursor is known as Type A and exhibits an isoelectric point between pH 7 and 9, while gelatin derived from an alkali-treated precursor is known as Type B and exhibits an isoelectric point between pH 4.7 and 5.2.

Gelatin for use in the manufacture of capsules in which to dispense medicines or for the coating of tablets may be colored with a certified color, may contain not more than 0.15% of sulfur dioxide, may contain a suitable concentration of sodium lauryl sulfate and suitable antimicrobial agents, and may have any suitable gel strength that is designated by Bloom Gelometer number.

Regarding the special gelatin for use in the preparation of emulsions, see Emulsions.

Description—Sheets, flakes, shreds, or a coarse-to-fine powder; faintly yellow or amber in color, the color varying in depth according to the particle size; slight, characteristic bouillon-like odor; stable in air when dry, but is subject to microbial decomposition when moist or in solution.

Solubility—Insoluble in cold water, but swells and softens when immersed in it, gradually absorbing from 5 to 10 times its own weight of water; soluble in hot water, acetic acid, or hot mixtures of glycerin or water; insoluble in alcohol, chloroform, ether, or fixed and volatile oils.

Uses—In pharmacy, to coat tablets and form capsules, and as a vehicle for suppositories. It also is recommended as an emulsifying agent. See under Emulsions in Chapters 20 and 39, also Suppositories; and Absorbable Gelatin Sponge. It also has been used as an adjuvant protein food in malnutrition.

GLYCERYL MONOSTEARATE—page 1078.

HYDROXYETHYL CELLULOSE

Cellulose, 2-hydroxyethyl ether; Cellosize; Natrosol

Cellulose hydroxyethyl ether 9004-62-0.

Preparation—Cellulose is treated with NaOH and then reacted with ethylene oxide.

Description—White, odorless, tasteless, free-flowing powder; softens at about 137°C; refractive index (2% solution) about 1.336; pH about 7; solutions are nonionic.

Solubility—Dissolves readily in cold or hot water to give clear, smooth, viscous solutions; partially soluble in acetic acid; insoluble in most organic solvents.

Uses—Resembles carboxymethylcellulose sodium in that it is a cellulose ether, but differs in being nonionic, and hence, its solutions are unaffected by cations. It is used pharmaceutically as a thickener, protective colloid, binder, stabilizer, and suspending agent in emulsions, jellies and ointments, lotions, ophthalmic solutions, suppositories, and tablets.

HYDROXYPROPYL CELLULOSE

Cellulose, 2-hydroxypropyl ether; Klucel

Cellulose hydroxypropyl ether [9004-64-2].

Preparation. elevated temperature and pressure.

Description—Off-white, odorless, tasteless powder; softens at 130°C; burns out.\ completely about 475°C in N_2 or O_2; refractive index (2% solution) about 1.337; pH (aqueous solution) 5 to 8.5; solutions are nonionic.

Solubility—Soluble in water below 40°C (insoluble above 45°C); soluble in many polar organic solvents.

Uses—A broad combination of properties useful in a variety of industries. It is used pharmaceutically as a binder, granulation agent, and film-coating in the manufacture of tablets; an alcohol-soluble thickener and suspending agent for elixirs and lotions; and a stabilizer for emulsions.

HYDROXYPROPYL METHYLCELLULOSE

Cellulose, 2-hydroxypropyl methyl ether

Cellulose hydroxypropyl methyl ether [9004-65-3], available in grades containing 16.5 to 30.0% of methoxy and 4.0 to 32.0% of hydroxypropoxy groups, and thus in viscosity and thermal gelation temperatures of solutions of specified concentration.

Preparation—The appropriate grade of methylcellulose (see below) is treated with NaOH and reacted with propylene oxide at elevated temperature and pressure for a reaction time sufficient to produce the desired degree of attachment of methyl and hydroxypropyl groups by ether linkages to the anhydroglucose rings of cellulose.

Description—White to slightly off-white, fibrous or granular, free-flowing powder.

Solubility—Swells in water and produces a clear to opalescent, viscous, colloidal mixture; undergoes reversible transformation from sol to gel on heating and cooling, respectively. Insoluble in anhydrous alcohol, ether, or chloroform.

Uses—A protective colloid that is useful as a dispersing and thickening agent, and in ophthalmic solutions to provide the demulcent action and viscous properties essential for contact-lens use and in artificial-tear formulations. Also used in the preparation of sustained release matrix tablets and as a film coating material.

LANOLIN, ANHYDROUS—page 1077.

METHYLCELLULOSE

Cellulose, methyl ether; Methocel

Cellulose methyl ether [9004-67-5]; a methyl ether of cellulose containing 27.5 to 31.5% of methoxy groups.

Preparation—By the reaction of methyl chloride or of dimethyl sulfate on cellulose dissolved in sodium hydroxide. The cellulose methyl ether so formed is coagulated by adding methanol or other suitable agent and centrifuged. Since cellulose has 3 hydroxyl groups/glucose residue, several methylcelluloses can be made that vary in, among other properties, solubility and viscosity. Types useful for pharmaceutical application contain from 1 to 2 methoxy radicals/glucose residue.

Description—White, fibrous powder or granules; aqueous suspensions neutral to litmus; stable to alkali and dilute acids.

Solubility—Insoluble in ether, alcohol, or chloroform; soluble in glacial acetic acid or in a mixture of equal parts of alcohol and chloroform; swells in water, producing a clear to opalescent, viscous colloidal solution; insoluble in hot water and saturated salt solutions; salts of minerals, acids, and particularly polybasic acids, phenols, and tannins coagulate its solutions, but this can be prevented by the addition of alcohol or of glycol diacetate.

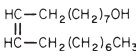
Uses—A synthetic substitute for natural gums that has both pharmaceutical and therapeutic applications. Pharmaceutically, it is used as a dispersing, thickening, emulsifying, sizing, and coating agent. It is an ingredient of many nose drops, eye preparations, burn medications, cosmetics, tooth pastes, liquid dentifrices, hair fixatives, creams, and lotions. It functions as a protective colloid for many types of dispersed substances and is an effective stabilizer for oil-in-water emulsions.

Therapeutically, it is used as a bulk laxative in the treatment of chronic constipation. Taken with 1 or 2 glassfuls of water, it forms a col-

loidal solution in the upper alimentary tract; this solution loses water in the colon, forming a gel that increases the bulk and softness of the stool. The gel is bland, demulcent, and nonirritating to the GI tract. Once a normal stool develops, the dose should be reduced to a level adequate for maintenance of good function. Although it takes up water from the GI tract quite readily, methylcellulose tablets have caused fecal impaction and intestinal obstruction when taken with a limited amount of water. It also is used as a topical ophthalmic protectant, in the form of 0.5 to 1% solution serving as artificial tears or a contact-lens solution applied to the conjunctiva, 0.05 to 0.1 mL at a time, 3 or 4 times a day as needed.

OLEYL ALCOHOL

9-Octadecen-1-ol, *(Z)*-, Aldol 85



(Z)-9-Octadecen-1-ol [143-28-2] $C_{18}H_{36}O$ (268.48); a mixture of unsaturated and saturated high-molecular-weight fatty alcohols consisting chiefly of oleyl alcohol.

Preparation—One method reacts ethyl oleate with absolute ethanol and metallic sodium (*Org Syn Coll III:* 673, 1955).

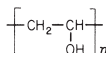
Description—Clear, colorless to light yellow, oily liquid; faint characteristic odor and bland taste; iodine value between 85 and 90; hydroxyl value between 205 and 215.

Solubility—Soluble in alcohol, ether, isopropyl alcohol, or light mineral oil; insoluble in water.

Uses—A *pharmaceutical aid* (emulsifying agent or emollient).

POLYVINYL ALCOHOL

Ethenol, homopolymer



Vinyl alcohol polymer [9002-89-5] $(C_2H_4O)_n$.

Preparation—Polyvinyl acetate is approximately 88% hydrolyzed in a methanol-methyl acetate solution using either mineral acid or alkali as a catalyst.

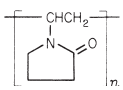
Description—White to cream-colored powder or granules; odorless.

Solubility—Freely soluble in water; solution effected more rapidly at somewhat elevated temperatures.

Uses—A suspending agent and emulsifier, either with or without the aid of a surfactant. It commonly is employed as a lubricant and protectant in various ophthalmic preparations, such as decongestants, artificial tears, and contact-lens products.

POVIDONE

2-Pyrrolidinone, 1-ethenyl-, homopolymer; Polyvinylpyrrolidone; PVP



1-Vinyl-2-pyrrolidinone polymer [9003-39-8] $(C_6H_9NO)_n$; a synthetic polymer consisting of linear 1-vinyl-2-pyrrolidinone groups, the degree of polymerization of which results in polymers of various molecular weights. It is produced commercially as a series of products having mean molecular weights ranging from about 10,000 to about 700,000. The viscosity of solutions containing 10% or less is essentially the same as that of water; solutions more concentrated than 10% become more viscous, depending upon the concentration and the molecular weight of the polymer used. It contains between 12 to 13% nitrogen.

Preparation—1,4-Butanediol is dehydrogenated thermally with the aid of copper to γ-butyrolactone, which then is reacted with ammonia to form 2-pyrrolidinone. Addition of the latter to acetylene yields vinylpyrrolidinone (monomer), which is polymerized thermally in the presence of hydrogen peroxide and ammonia.

Description—White to creamy white, odorless powder, hygroscopic; pH (1 in 20 solution) 3 to 7.

Solubility—Soluble in water, alcohol, or chloroform; insoluble in ether.

Uses—A dispersing and suspending agent in pharmaceutical preparations. Also used as a binder in wet granulation processes.

PROPYLENE GLYCOL MONOSTEARATE

Octadecanoic acid, monoester with 1,2-propanediol

1,2-Propanediol monostearate [1323-39-3]; a mixture of the propylene glycol mono- and diesters of stearic and palmitic acids. It contains not less than 90% monoesters of saturated fatty acids, chiefly propylene glycol monostearate ($C_{21}H_{42}O_3$) and propylene glycol monopalmitate ($C_{19}H_{38}O_3$).

Preparation—By reacting propylene glycol with stearoyl chloride in a suitable dehydrochlorinating environment.

Description—White, wax-like solid or white, wax-like beads or flakes; slight, agreeable, fatty odor and taste; congeals not lower than 45°C; acid value not more than 2; saponification value 155 to 165; hydroxyl value 150 to 170; iodine value not more than 3.

Solubility—Dissolves in organic solvents such as alcohol, mineral or fixed oils, benzene, ether, or acetone; insoluble in water but may be dispersed in hot water with the aid of a small amount of soap or other suitable surface-active agent.

Uses—A surfactant. It is particularly useful as a dispersing agent for perfume oils or oil-soluble vitamins in water, and in cosmetic preparations.

SILICON DIOXIDE, COLLOIDAL—page 1089.

SODIUM LAURYL SULFATE

Sulfuric acid monododecyl ester sodium salt; Irium; Duponol C; Gardinol WA

Sodium monododecyl sulfate [151-21-3]; a mixture of sodium alkyl sulfates consisting chiefly of sodium lauryl sulfate. The combined content of sodium chloride and sodium sulfate is not more than 8%.

Preparation—The fatty acids of coconut oil, consisting chiefly of lauric acid, are catalytically hydrogenated to form the corresponding alcohols. The latter are then esterified with sulfuric acid (sulfated) and the resulting mixture of alkyl bisulfates (alkylsulfuric acids) is converted into a mixture of sodium salts by reacting with alkali under controlled conditions of pH.

Description—Small, white or light yellow crystals having a slight, characteristic odor.

Solubility—1 g in 10 mL water, forming an opalescent solution.

Incompatibilities—Reacts with cationic surface-active agents with loss of activity, even in concentrations too low to cause precipitation. Unlike soaps, it is compatible with dilute acids and calcium and magnesium ions.

Uses—An emulsifying, detergent, and wetting agent in ointments, tooth powders, and other pharmaceutical preparations, and in the metal, paper, and pigment industries.

SORBITAN ESTERS

Spans

Sorbitan esters (monolaurate [1338-39-2]; monooleate [1338-43-8]; monopalmitate [26266-57-9]; monostearate [1338-41-6]; trioleate [26266-58-0]; tristearate [26658-19-5]).

Preparation—Sorbitol is dehydrated to form a hexitan that is then esterified with the desired fatty acid which are polyethylene glycol ethers of sorbitan fatty acid esters.

Description—Monolaurate: Amber, oily liquid; may become hazy or form a precipitate; viscosity about 4250 cps; HLB number 8.6; acid number 7.0 max; saponification number 158 to 170; hydroxyl number 330 to 358. Monooleate: Amber liquid; viscosity about 1000 cps; HLB number 4.3; acid number 8.0 max; saponification number 145 to 160; hydroxyl number 193 to 210. Monopalmitate: Tan, granular waxy solid; HLB number 6.7; acid number 4 to 7.5; saponification number 140 to 150; hydroxyl number 275 to 305. Monostearate: Cream to tan beads; HLB number 4.7; acid number 5 to 10; saponification number 147 to 157; hydroxyl number 235 to 260. Trioleate: Amber, oily liquid; viscosity about 200 cps; HLB number 1.8; acid number 15 max; saponification number 170 to 190; hydroxyl number 55 to 70. Tristearate: Tan, waxy beads; HLB number 2.1; acid number 12 to 15; saponification number 176 to 188; hydroxyl number 66 to 80.

Solubility—Monolaurate: Soluble in methanol or alcohol; dispersible in distilled water and hard water (200 ppm); insoluble in hard water (20,000 ppm). Monooleate: Soluble in most mineral or vegetable oils; slightly soluble in ether; dispersible in water; insoluble in acetone. Monopalmitate: Dispersible in distilled water or hard water (200 ppm); soluble in ethyl acetate; insoluble in cold distilled water or hard water (20,000 ppm). Monostearate: Soluble (above melting point) in vegetable oils or mineral oil; insoluble in water, alcohol, or propylene glycol. Trioleate: Soluble in mineral oil, vegetable oils, alcohol, or

methanol; insoluble in water. Tristearate: Soluble in isopropyl alcohol; insoluble in water.

Uses—Nonionic surfactants used as emulsifying agents in the preparation of water-in-oil emulsions.

STEARIC ACID—page 1079.

STEARYL ALCOHOL

1-Octadecanol [112-92-5] $C_{18}H_{38}O$ (270.50); contains not less than 90% of stearyl alcohol, the remainder consisting chiefly of cetyl alcohol [$C_{16}H_{34}O$ = 242.44].

Preparation—Through the reducing action of lithium aluminum hydride on ethyl stearate.

Description—White, unctuous flakes or granules having a faint, characteristic odor and a bland taste; melts 55 to 60°C.

Solubility—Insoluble in water; soluble in alcohol, chloroform, ether, or vegetable oils.

Uses—A surface-active agent used to stabilize emulsions and increase their ability to retain large quantities of water. See Hydrophilic Ointment (page 1078); Hydrophilic Petrolatum (page 1078).

TRAGACANTH

Gum Tragacanth; Hog Gum; Goat's Thorn

The dried gummy exudation from *Astragalus gummifer* Labillardiére or other Asiatic species of *Astragalus* (Fam *Leguminosae*).

Constituents—60 to 70% bassorin and 30 to 40% soluble gum *(tragacanthin)*. The bassorin swells in the presence of water to form a gel, and tragacanthin forms a colloidal solution. Bassorin, consisting of complex methoxylated acids, resembles pectin. Tragacanthin yields glucuronic acid and arabinose when hydrolyzed.

Description—Flattened, lamellated, frequently curved fragments or straight or spirally twisted linear pieces 0.5 to 2.5 mm in thickness; white to weak-yellow in color; translucent; horny in texture; odorless; insipid, mucilaginous taste. When powdered, it is white to yellowish white.

Introduced into water, tragacanth absorbs a certain proportion of that liquid, swells very much, forms a soft adhesive paste, but does not dissolve. If agitated with an excess of water, this paste forms a uniform mixture; but in the course of 1 or 2 days the greater part separates and is deposited, leaving a portion dissolved in the supernatant fluid. The finest mucilage is obtained from the whole gum or flake form. Several days should be allowed for obtaining a uniform mucilage of the maximum gel strength. A common adulterant is Karaya Gum, and the USP has introduced tests to detect its presence.

Solubility—Insoluble in alcohol.

Uses—A suspending agent in lotions, mixtures, and extemporaneous preparations and prescriptions. It is used with emulsifying agents largely to increase consistency and retard creaming. It is sometimes used as a demulcent in sore throat, and the jelly-like product formed when the gum is allowed to swell in water serves as a basis for pharmaceutical jellies, eg, Ephedrine Sulfate Jelly. It also is used in various confectionery products. In the form of a glycerite, it has been used as a pill excipient.

Tragacanth Mucilage—Preparation: Mix glycerin (18 g) with purified water (75 mL) in a tared vessel, heat the mixture to boiling, discontinue the application of heat, add tragacanth (6 g) and benzoic acid (0.2 g), and macerate the mixture during 24 hr, stirring occasionally. Then add enough purified water to make the mixture weigh 100 g, stir actively until of uniform consistency, and strain forcibly through muslin. Uses: A suspending agent for insoluble substances in internal mixtures. It is also a protective agent.

XANTHAN GUM

Keltrol

A high-molecular-weight polysaccharide gum produced by a pure- culture fermentation of a carbohydrate with *Xanthomonas campestris,* then purified by recovery with isopropyl alcohol, dried and milled; contains D-glucose and D-mannose as the dominant hexose units, along with D-glucuronic acid and is prepared as a sodium, potassium, or calcium salt; yields 4.2 to 5% carbon dioxide.

Preparation—See above and US Patents 3,433,708 and 3,557,016.

Description—White or cream-colored, tasteless powder with a slight organic odor; powder and solutions stable at 25°C or less; does not exhibit polymorphism; aqueous solutions are neutral to litmus.

Solubility—1 g in about 3 mL alcohol; soluble in hot or cold water.

Uses—A hydrophilic colloid to thicken, suspend, emulsify, and stabilize water-based systems.

OTHER EMULSIFYING AND SUSPENDING AGENTS

Malt—The partially germinated grain of one or more varieties of *Hordeum vulgare* Linné (Fam *Gramineae*) and contains amylolytic enzymes. Yellowish or amber-colored grains, with a characteristic odor and a sweet taste. The evaporated aqueous extract constitutes malt extract.

Malt Extract—The product obtained by extracting malt, the partially and artificially germinated grain of one or more varieties of *Hordeum vulgare* Linné (Fam *Gramineae*). Uses: An infrequently used emulsifying agent.

OINTMENT BASES

Ointments are semisolid preparations for external application to the body. They should be of such composition that they soften, but not necessarily melt, when applied to the skin. Therapeutically, ointments function as protectives and emollients for the skin, but are used primarily as vehicles or bases for the topical application of medicinal substances. Ointments also may be applied to the eye or eyelids.

Ideally, an ointment base should be compatible with the skin, stable, permanent, smooth and pliable, nonirritating, nonsensitizing, inert, and readily able to release its incorporated medication. Since there is no single ointment base that possesses all these characteristics, continued research in this field has resulted in the development of numerous new bases. Indeed, ointment bases have become so numerous as to require classification. Although ointment bases may be grouped in several ways, it is generally agreed that they can be classified best according to composition. Hence, the following four classes are recognized here: oleaginous, emulsifiable, emulsion bases, and water-soluble.

For completeness, substances are included that, although not used alone as ointment bases, contribute some pharmaceutical property to one or more of the various bases.

Oleaginous Ointment Base and Components

The oleaginous ointment bases include fixed oils of vegetable origin, fats obtained from animals, and semisolid hydrocarbons obtained from petroleum. The vegetable oils are used chiefly in ointments to lower the melting point or to soften bases. These oils can be used as a base in themselves when a high percentage of powder is incorporated.

The vegetable oils and the animal fats have two marked disadvantages as ointment bases: their water-absorbing capacity is low and they have a tendency to become rancid. Insofar as vegetable oils are concerned, the second disadvantage can be overcome by hydrogenation, a process that converts many fixed oils into white, semisolid fats or hard, almost brittle, waxes.

The hydrocarbon bases comprise a group of substances with a wide range of melting points so that any desired consistency and melting point may be prepared with representatives of this group. They are stable, bland, and chemically inert and will mix with virtually any chemical substance. Oleaginous bases are excellent emollients.

WHITE OINTMENT

Ointment USP; Simple Ointment

White Wax	50 g
White Petrolatum	950 g
To make	1000 g

Melt the white wax in a suitable dish on a water bath, add the white petrolatum, warm until liquefied, then discontinue the heating and stir the mixture until it begins to congeal. It is permissible to vary the proportion of wax to obtain a suitable consistency of the ointment under different climatic conditions.

Uses—An emollient and vehicle for other ointments.

YELLOW OINTMENT

Yellow Wax	50 g
Petrolatum	950 g
To make	1000 g

Melt the yellow wax in a suitable dish on a steam bath, add the petrolatum, warm until liquefied, then discontinue the heating and stir the mixture until it begins to congeal. It is permissible to vary the proportion of wax to obtain a suitable consistency of the ointment under different climatic conditions.

Uses—An emollient and vehicle for other ointments. Both white and yellow ointments are known as simple ointment. White ointment should be used to prepare white ointments and yellow ointments should be used to prepare colored ointments when simple ointment is prescribed.

CETYL ESTERS WAX

Synthetic Spermaceti

A mixture consisting primarily of esters of saturated fatty alcohols (C_{14} to C_{18}) and saturated fatty acids (C_{14} to C_{18}). It has a saponification value of 109 to 120 and an acid value of not more than 5.

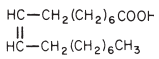
Description—White to off-white, somewhat translucent flakes; crystalline structure and pearly luster when caked; faint odor and a bland, mild taste; free from rancidity; specific gravity 0.820 to 0.840 at 50°C; iodine value not more than 1; melts 43 to 47°C.

Solubility—Insoluble in water; practically insoluble in cold alcohol; soluble in boiling alcohol, ether, chloroform, or fixed and volatile oils; slightly soluble in cold solvent hexane.

Uses—A replacement for spermaceti used to give consistency and texture to ointments, eg, Cold Cream and Rose Water Ointment.

OLEIC ACID

***(Z)*-9-Octadecenoic acid; Oleinic Acid; Elaic Acid**



Oleic acid [112-80-1] obtained from tallow and other fats and consists chiefly of *(Z)*-9-octadecenoic acid (282.47). Oleic acid used in preparations for internal administration is derived from edible sources.

It usually contains variable amounts of the other fatty acids present in tallow, such as linolenic and stearic acids.

Preparation—Obtained as a by-product in the manufacture of the solid stearic and palmitic acids used in the manufacture of candles, stearates, and other products. The crude oleic acid is known as red oil, the stearic and palmitic acids being separated by cooling.

Description—Colorless to pale yellow, oily liquid; lard-like odor and taste; specific gravity 0.889 to 0.895; congeals at a temperature not above 10°C; pure acid solidifies at 4°C; at atmospheric pressure it decomposes when heated at 80 to 100°C; on exposure to air it gradually absorbs oxygen, darkens, and develops a rancid odor.

Solubility—Practically insoluble in water; miscible with alcohol, chloroform, ether, benzene, or fixed and volatile oils.

Incompatibilities—Reacts with alkali to form soaps. Heavy metals and calcium salts form insoluble oleates. Iodine solutions are decolorized by formation of the iodine addition compound of the acid. It is oxidized to various derivatives by nitric acid, potassium permanganate, and other agents.

Uses—Classified as an emulsion adjunct, which reacts with alkalis to form soaps that function as emulsifying agents; it is used for this purpose in such preparations as Benzyl Benzoate Lotion and Green Soap. It also is used to prepare oleate salts of bases.

PARAFFIN

Paraffin Wax; Hard Paraffin

A purified mixture of solid hydrocarbons obtained from petroleum.

Description—Colorless or white, more or less translucent mass with a crystalline structure; slightly greasy to the touch; odorless and tasteless; congeals 47 to 65°C.

Solubility—Freely soluble in chloroform, ether, volatile oils, or most warm fixed oils; slightly soluble in dehydrated alcohol; insoluble in water or alcohol.

Uses—Mainly, to increase the consistency of some ointments.

PETROLATUM

Yellow Soft Paraffin; Amber Petrolatum; Yellow Petrolatum; Petroleum Jelly; Paraffin Jelly

A purified mixture of semisolid hydrocarbons obtained from petroleum. It may contain a suitable stabilizer.

Preparation—The residuums as they are termed technically, which are obtained by the distillation of petroleum, are purified by melting, usually treating with sulfuric acid and then percolating through recently burned bone black or adsorptive clays; this removes the odor and modifies the color. Selective solvents are also sometimes employed to extract impurities.

It has been found that the extent of purification required to produce Petrolatum and Light Mineral Oil of official quality removes antioxidants that are naturally present, and the purified product subsequently has a tendency to oxidize and develop an offensive odor. This is prevented by the addition of a minute quantity of α-tocopherol or other suitable antioxidant, as is now permissible.

Description—Unctuous mass of yellowish to light amber color; not more than a slight fluorescence after being melted; transparent in thin layers; free or nearly free from odor and taste; specific gravity 0.815 to 0.880 at 60°C; melts between 38 and 60°C.

Solubility—Insoluble in water; almost insoluble in cold or hot alcohol or in cold dehydrated alcohol; freely soluble in benzene, carbon disulfide, chloroform, or turpentine oil; soluble in ether, solvent hexane, or in most fixed and volatile oils, the degree of solubility in these solvents varying with the composition of the petrolatum.

Uses—A base for ointments. It is highly occlusive and therefore a good emollient, but it may not release some drugs readily.

WHITE PETROLATUM

White Petroleum Jelly; White Soft Paraffin

A purified mixture of semisolid hydrocarbons obtained from petroleum, and wholly or nearly decolorized. It may contain a suitable stabilizer.

Preparation—In the same manner as petrolatum, the purification treatment being continued until the product is practically free from yellow color.

Description—White or faintly yellowish, unctuous mass; transparent in thin layers, even after cooling to 0°C; specific gravity 0.815 to 0.880 at 60°C; melts 38 to 60°C.

Solubility—Similar to that described under Petrolatum.

Uses—Similar to yellow petrolatum but often is preferred because of its freedom from color. It is employed as a protective, as a base for ointments and cerates, and to form the basis for burn dressings. See Petrolatum Gauze (page 1278).

Absorbent Ointment Bases

The term absorbent is used here to denote the water-absorbing or emulsifying properties of these bases and not to describe their action on the skin. These bases, sometimes called emulsifiable ointment bases, are generally anhydrous substances that have the property of absorbing (emulsifying) considerable quantities of water and still retaining their ointment-like consistency. Preparations of this type do not contain water as a component of their basic formula, but if water is incorporated, when and as desired, a W/O emulsion results. The following official products fall into this category.

LANOLIN ANHYDROUS

Anhydrous Lanolin; Wool Fat USP; Refined Wool Fat

Lanolin that contains not more than 0.25% of water.

Constituents—Contains the sterols cholesterol [$C_{27}H_{45}OH$] and oxycholesterol, as well as triterpene and aliphatic alcohols. About 7% of the alcohols are found in the free state, the remainder occurring as es-

ters of the following fatty acids: carnaubic, cerotic, lanoceric, lanopalmitic, myristic, and palmitic. Some of these are found free. The emulsifying and emollient actions of lanolin are due to the alcohols that are found in the unsaponifiable fraction when lanolin is treated with alkali. Constituting approximately one-half of this fraction and known as lanolin alcohols, the latter is composed of cholesterol (30%), lanosterol (25%), cholestanol (dihydrocholesterol) (3%), agnosterol (2%), and various other alcohols (40%).

Preparation—By purifying the fatty matter (suint) obtained from the wool of the sheep. This natural wool fat contains about 30% of free fatty acids and fatty acid esters of cholesterol and other higher alcohols. The cholesterol compounds are the important constituents, and to secure these in a purified form, many processes have been devised. In one of these the crude wool fat is treated with weak alkali and the saponified fats and emulsions are centrifuged to secure the aqueous soap solution, from which, on standing, a layer of partially purified wool fat separates. This product is further purified by treating it with calcium chloride and then dehydrated by fusion with unslaked lime. It is finally extracted with acetone, and the solvent subsequently separated by distillation. This differs from lanolin in that the former contains practically no water.

Description—Yellow, tenacious, unctuous mass; slight, characteristic odor; melts between 36 and 42°C.

Solubility—Insoluble in water but mixes without separation with about twice its weight of water; sparingly soluble in cold alcohol; more soluble in hot alcohol; freely soluble in ether or chloroform.

Uses—An ingredient of ointments, especially when an aqueous liquid is to be incorporated. It gives a distinctive quality to the ointment, increasing absorption of active ingredients and maintaining a uniform consistency for the ointment under most climatic conditions. However, it has been omitted from many ointments on the recommendation of dermatologists who have found that many patients are allergic to this animal wax.

HYDROPHILIC PETROLATUM

Cholesterol	30 g
Stearyl Alcohol	30 g
White Wax	80 g
White Petrolatum	860 g
To make	1000 g

Melt the stearyl alcohol, white wax, and white petrolatum together on a steam bath, then add the cholesterol and stir until it completely dissolves. Remove from the bath, and stir until the mixture congeals.

Uses—A protective and water-absorbable ointment base. It will absorb a large amount of water from aqueous solutions of medicating substances, forming a W/O type of emulsion. See Chapter 44 *(Medicated Topicals)*.

Emulsion Ointment Bases and Components

Emulsion ointment bases are actually semisolid emulsions. These preparations can be divided into two groups on the basis of emulsion type: emulsion ointment base water-in-oil (W/O) type and emulsion ointment base oil-in-water (O/W) type. Bases of both types will permit the incorporation of some additional amounts of water without reducing the consistency of the base below that of a soft cream. However, only O/W emulsion ointment bases can be removed readily from the skin and clothing with water. W/O emulsions are better emollients and protectants than are O/W emulsions. W/O emulsions can be diluted with oils.

CETYL ALCOHOL

Cetostearyl Alcohol; Palmityl Alcohol; Aldol 52

$CH_3(CH_2)_{14}CH_2OH$

1-Hexadecanol [124-29-8] $C_{16}H_{34}O$ (242.44); a mixture of not less than 90% of cetyl alcohol, the remainder chiefly stearyl alcohol.

Preparation—By catalytic hydrogenation of palmitic acid or saponification of spermaceti, which contains cetyl palmitate.

Description—Unctuous, white flakes, granules, cubes, or castings; faint characteristic odor and a bland, mild taste; melts 45 to 50°C; not less than 90% distills between 316 and 336°C.

Solubility—Insoluble in water; soluble in alcohol, chloroform, ether, or vegetable oils.

Uses—Similar to Stearyl Alcohol (page 1076). It also imparts a smooth texture to the skin and is used widely in cosmetic creams and lotions.

COLD CREAM

Petrolatum Rose Water Ointment USP

Cetyl Esters Wax	125 g
White Wax	120 g
Mineral Oil	560 g
Sodium Borate	5 g
Purified Water	190 mL
To make about	1000 g

Reduce the cetyl esters wax and the white wax to small pieces, melt them on a steam bath with the mineral oil, and continue heating until the temperature of the mixture reaches 70°C. Dissolve the sodium borate in the purified water, warmed to 70°C, and gradually add the warm solution to the melted mixture, stirring rapidly and continuously until it has congealed.

If the ointment has been chilled, warm it slightly before attempting to incorporate other ingredients (see USP for allowable variations).

Uses—Useful as an emollient, cleansing cream, and ointment base. It resembles *Rose Water Ointment,* differing only in that mineral oil is used in place of almond oil and omitting the fragrance. This change produces an ointment base that is not subject to rancidity as is one containing a vegetable oil. This is a W/O emulsion.

GLYCERYL MONOSTEARATE

Octadecanoic acid, monoester with 1,2,3-propanetriol

Monostearin [31566-31-1]; a mixture chiefly of variable proportions of glyceryl monostearate $[C_3H_5(OH)_2C_{18}H_{35}O_2 = 358.56]$ and glyceryl monopalmitate $[C_3H_5(OH)_2C_{16}H_{31}O_2 = 330.51]$.

Preparation—Among other ways, by reacting glycerin with commercial stearoyl chloride.

Description—White, wax-like solid or occurs in the form of white, wax-like beads, or flakes; slight, agreeable, fatty odor and taste; does not melt below 55°C; affected by light.

Solubility—Insoluble in water, but may be dispersed in hot water with the aid of a small amount of soap or other suitable surface-active agent; dissolves in hot organic solvents such as alcohol, mineral or fixed oils, benzene, ether, or acetone.

Uses—A thickening and emulsifying agent for ointments. See *Ointments* (page 1076).

HYDROPHILIC OINTMENT

Methylparaben	0.25 g
Propylparaben	0.15 g
Sodium Lauryl Sulfate	10 g
Propylene Glycol	120 g
Stearyl Alcohol	250 g
White Petrolatum	250 g
Purified Water (qs)	1000 g

Melt the stearyl alcohol and the white petrolatum on a steam bath, and warm to about 75°C. Add the other ingredients, previously dissolved in the water and warmed to 75°C, and stir the mixture until it congeals.

Uses A water-removable ointment base for the so-called washable ointments. This is an O/W emulsion.

LANOLIN

Hydrous Wool Fat

The purified, fat-like substance from the wool of sheep, *Ovis aries* Linné (Fam *Bovidae*); contains 25% to 30% water.

Description—Yellowish white, ointment-like mass, with a slight, characteristic odor; when heated on a steam bath it separates into an upper oily and a lower water layer; when the water is evaporated a residue of *Lanolin* remains that is transparent when melted.

Solubility—Insoluble in water; soluble in chloroform or ether with separation of its water of hydration.

Uses—Largely as a vehicle for ointments, for which it is admirably adapted on account of its compatibility with skin lipids. It emulsifies aqueous liquids. Lanolin is a W/O emulsion.

ROSE WATER OINTMENT

Cold Cream; Galen's Cerate

Cetyl Esters Wax	125 g
White Wax	120 g
Almond Oil	560 g
Sodium Borate	5 g
Stronger Rose Water	25 mL
Rose Oil	0.2 mL
Purified Water (qs)	1000 g

Reduce the cetyl esters wax and the white wax to small pieces, melt them on a steam bath, add the almond oil, and continue heating until the temperature of the mixture reaches 70°C. Dissolve the sodium borate in the purified water and stronger rose water, warmed to 70°C, and gradually add the warm solution to the melted mixture, stirring rapidly and continuously until it has cooled to about 45°C. Incorporate the rose oil.

It must be free from rancidity. If the ointment has been chilled, warm it slightly before attempting to incorporate other ingredients (see USP for allowable variations).

History—Originated by Galen, the famous Roman physician- pharmacist of the 1st Century AD; was known for many centuries by the name of *Unguentum* or *Ceratum Refrigerans*. It has changed but little in proportions or method of preparation throughout many centuries.

Uses—An emollient and ointment base. It is a W/O emulsion.

STEARIC ACID

Octadecanoic acid; Cetylacetic Acid; Stearophanic Acid

Stearic acid [57-11-4]; a mixture of stearic acid [$C_{18}H_{36}O_2$ = 284.48] and palmitic acid [$C_{16}H_{32}O_2$ = 256.43], which together constitute not less than 90.0% of the total content. The content of each is not less than 40.0% of the total.

Purified Stearic Acid USP is a mixture of the same acids that together constitute not less than 96.0% of the total content, and the content of $C_{18}H_{36}O_2$ is not less than 90.0% of the total.

Preparation—From edible fats and oils (see exception below) by boiling them with soda lye, separating the glycerin, and decomposing the resulting soap with sulfuric or hydrochloric acid. The stearic acid subsequently is separated from any oleic acid by cold expression. It also is prepared by the hydrogenation and subsequent saponification of olein. It may be purified by recrystallization from alcohol.

Description—Hard, white or faintly yellowish, somewhat glossy and crystalline solid, or a white or yellowish white powder; an odor and taste suggestive of tallow; melts about 55.5°C and should not congeal at a temperature below 54°C; the purified acid melts at 69 to 70°C and congeals between 66 and 69°C; slowly volatilizes between 90 and 100°C.

Solubility—Practically insoluble in water; 1 g in about 20 mL alcohol, 2 mL chloroform, 3 mL ether, 25 mL acetone, or 6 mL carbon tetrachloride; freely soluble in carbon disulfide; also soluble in amyl acetate, benzene, or toluene.

Incompatibilities—Insoluble stearates are formed with many *metals*. Ointment bases made with stearic acid may show evidence of drying out or lumpiness due to such a reaction when zinc or calcium salts are compounded therein.

Uses—In the preparation of sodium stearate, which is the solidifying agent for the official glycerin suppositories; in enteric tablet coating; ointments; and for many other commercial products, such as toilet creams, vanishing creams, solidified alcohol, etc. (when labeled solely for external use, it is exempt from the requirement that it be prepared from edible fats and oils).

Water-Soluble Ointment Bases and Components

Included in this section are bases prepared from the higher ethylene glycol polymers (PEGs). These polymers are marketed under the trademark of Carbowax. The polymers have a wide range in molecular weight. Those with molecular weights ranging from 200 to 700 are liquids; those above 1000 are wax-like solids. The polymers are water-soluble, nonvolatile, and unctuous agents. They do not hydrolyze or deteriorate and will not support mold growth. These properties account for their wide use in washable ointments. Mixtures of PEGs are used to give bases of various consistency, such as very soft to hard bases for suppositories.

GLYCOL ETHERS AND DERIVATIVES

This special class of ethers is of considerable importance in pharmaceutical technology. Both mono- and polyfunctional compounds are represented in the group. The simplest member is ethylene oxide, [—CH_2CH_2—O], the internal or cyclic ether of the simplest glycol, ethylene glycol [$HOCH_2CH_2OH$]. External mono- and diethers of ethylene glycol $ROCH_2CH_2OH$ and $ROCH_2CH_2OR$ are well known largely because of research done by Union Carbide.

PREPARATION—In the presence of NaOH at temperatures of the order of 120 to 135°C and under a total pressure of about 4 atmospheres, ethylene oxide reacts with ethylene glycol to form compounds having the general formula $HOCH_2(CH_2OCH_2)_nCH_2OH$, commonly referred to as condensation polymers and termed polyethylene (or polyoxyethylene) glycols. Other glycols besides ethylene glycol function in a similar capacity, and the commercial generic term adopted for the entire group is polyalkylene (or polyoxyalkylene) glycols.

NOMENCLATURE—It is to be noted that these condensation polymers are bifunctional; ie, they contain both ether and alcohol linkages. The compound in which n = 1 is the commercially important diethylene glycol [$HOCH_2CH_2OCH_2CH_2OH$], and its internal ether is the familiar dioxane [—$CH_2CH_2OCH_2CH_2$—O]. The mono- and diethers derived from diethylene glycol have the formulas $ROCH_2CH_2OCH_2CH_2OH$ and $ROCH_2CH_2OCH_2CH_2OR$. The former is commonly termed Carbitols and the latter Cellosolves.

Polyethylene glycols are differentiated in commercial nomenclature by adding a number to the name, which represents the average molecular weight. Thus, polyethylene glycol 400 has an average molecular weight of about 400 (measured values for commercial samples range between 380 and 420), corresponding to a value of n for this particular polymer of approximately 8. Polymers have been produced in which the value of n runs into the hundreds. Up to n = approximately 15, the compounds are liquids at room temperature, and viscosity and boiling point increase with increasing molecular weight. Higher polymers are waxy solids and are termed commercially Carbowaxes.

It should be observed that the presence of the two terminal hydroxyl groups in the polyalkylene glycols makes possible the formation of both ether and ester derivatives, several of which are marketed products.

USES—Because of their vapor pressure, solubility, solvent power, hygroscopicity, viscosity, and lubricating characteristics, the polyalkylene glycols or their derivatives function in many applications as effective replacements for glycerin and water-insoluble oils. They find considerable use as plasticizers, lubricants, conditioners, and finishing agents for processing textiles and rubber. They also are important as emulsifying agents and as dispersants for such diverse substances as dyes, oils, resins, insecticides, and various types of pharmaceuticals. In addition, they are employed frequently as ingredients in ointment bases and in a variety of cosmetic preparations.

POLYETHYLENE GLYCOLS

Poly(oxy-1,2-ethanediyl), α-hydro-φ-hydroxy-, Carbowaxes; Atpeg

$$H—[OCH_2CH_2—]_nOH$$

Polyethylene glycols [25322-68-3].

Preparation—Ethylene glycol is reacted with ethylene oxide in the presence of NaOH at temperatures in the range of 120 to 135°C under pressure of about 4 atm.

Description—Polyethylene glycols 200, 300, 400, and 600 are clear, viscous liquids at room temperature. Polyethylene glycols 900, 1000, 1450, 3350, 4500, and 8000 are white, waxy solids. The glycols do not hydrolyze or deteriorate under typical conditions. As their molecular weight increases, their water solubility, vapor pressure, hygroscopicity, and solubility in organic solvents decrease; at the same time, freezing or melting range, specific gravity, flash point, and viscosity increase. If these compounds ignite, small fires should be extinguished with carbon dioxide or dry-chemical extinguishers and large fires with alcohol-type foam extinguishers.

Solubility—All members of this class dissolve in water to form clear solutions and are soluble in many organic solvents.

Uses—These possess a wide range of solubilities and compatibilities, which make them useful in pharmaceutical and cosmetic preparations. Their blandness renders them highly acceptable for hair dressings, hand lotions, sun-tan creams, leg lotions, shaving creams, and

skin creams (eg, a peroxide ointment that is stable may be prepared using these compounds, while oil-type bases inactivate the peroxide). Their use in washable ointments is discussed under Ointments (page 1076). They also are used in making suppositories, hormone creams, etc. See Polyethylene Glycol Ointment (below) and Glycol Ethers (above). The liquid polyethylene glycol 400 and the solid polyethylene glycol 3350, used in the proportion specified (or a permissible variation thereof) in the official Polyethylene Glycol Ointment, provide a water-soluble ointment base used in the formulation of many dermatological preparations. The solid, waxy, water-soluble glycols often are used to increase the viscosity of liquid polyethylene glycols and to stiffen ointment and suppository bases. In addition, they are used to compensate for the melting point-lowering effect of other agents, ie, chloral hydrate, etc, on such bases.

Polyethylene Glycol Ointment USP—Preparation: Heat polyethylene glycol 3350 (400 g) and polyethylene glycol 400 (600 g) on a water bath to 65°C. Allow to cool, and stir until congealed. If a firmer preparation is desired, replace up to 100 g of polyethylene glycol 400 with an equal amount of polyethylene glycol 3350. If 6 to 25% of an aqueous solution is to be incorporated in this ointment, replace 50 g of polyethylene glycol 3350 by 50 g of stearyl alcohol. Uses: A water-soluble ointment base.

POLYOXYL 40 STEARATE

Poly(oxy-1,2-ethanediyl), α-hydro-φ-hydroxy-, octadecanoate; Myrj

RCOO(C_2H_4O)$_n$H (RCOO is the stearate moiety; n is approximately 40).

Polyethylene glycol monostearate [9004-99-3]; a mixture of monostearate and distearate esters of mixed polyoxyethylene diols and corresponding free glycols, the average polymer length being equivalent to about 40 oxyethylene units. Polyoxyethylene 50 Stearate is a similar mixture in which the average polymer length is equivalent to about 50 oxyethylene units.

Preparation—One method consists of heating the corresponding polyethylene glycol with an equimolar portion of stearic acid.

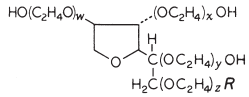
Description—White to light-tan waxy solid; odorless or has a faint fat-like odor; congeals between 37 and 47°C.

Solubility—Soluble in water, alcohol, ether, or acetone; insoluble in mineral or vegetable oils.

Uses—Contains ester and alcohol functions that impart both lyophilic and hydrophilic characteristics to make it useful as a surfactant and emulsifier. It is an ingredient of some water-soluble ointment and cream bases.

POLYSORBATES

Sorbitan esters, poly(oxy-1,2-ethanediyl) derivs; Tweens



[Sum of $w, x, y,$ and z is 20; R is $(C_{11}H_{23})COO$]

Sorbitan esters, polyoxyethylene derivatives; fatty acid esters of sorbitol and its anhydrides copolymerized with a varying number of moles of ethylene oxide. The NF recognizes Polysorbate 20 (structure given above), a laurate ester; Polysorbate 40, a palmitate ester; Polysorbate 60, a mixture of stearate and palmitate esters; and Polysorbate 80, an oleate ester.

Preparation—These important nonionic surfactants are prepared starting with sorbitol by (1) elimination of water-forming sorbitan (a cyclic sorbitol anhydride); (2) partial esterification of the sorbitan with a fatty acid such as oleic or stearic acid, yielding a hexitan ester known commercially as a Span; and (3) chemical addition of ethylene oxide, yielding a Tween (the polyoxyethylene derivative).

Description—Polysorbate 80: Lemon- to amber-colored, oily liquid; faint, characteristic odor; warm, somewhat bitter taste; specific gravity 1.07 to 1.09; pH (1:20 aqueous solution) 6 to 8.

Solubility—Polysorbate 80: Very soluble in water, producing an odorless and nearly colorless solution; soluble in alcohol, cottonseed oil, corn oil, ethyl acetate, methanol, or toluene; insoluble in mineral oil.

Uses—Because of their hydrophilic and lyophilic characteristics, these nonionic surfactants are very useful as emulsifying agents, forming O/W emulsions in pharmaceuticals, cosmetics, and other types of products. Polysorbate 80 is an ingredient in Coal Tar Ointment and Solution. See Glycol Ethers (page 1079).

PHARMACEUTICAL SOLVENTS

The remarkable growth of the solvent industry is attested by the more than 300 solvents now being produced on an industrial scale. Chemically, these include a great variety of organic compounds, ranging from hydrocarbons through alcohols, esters, ethers, and acids to nitroparaffins. Their main applications are in industry and the synthesis of organic chemicals. Comparatively few, however, are used as solvents in pharmacy, because of their toxicity, volatility, instability, and/or flammability. Those commonly used as pharmaceutical solvents are described in this section.

ACETONE

2-Propanone; Dimethyl Ketone

CH_3COCH_3

Acetone [67-64-1] C_3H_6O (58.08).

Caution—It is very flammable. Do not use where it may be ignited.

Preparation—Formerly obtained exclusively from the destructive distillation of wood. The distillate, consisting principally of methanol, acetic acid, and acetone was neutralized with lime, and the acetone was separated from the methyl alcohol by fractional distillation. Additional quantities were obtained by pyrolysis of the calcium acetate formed in the neutralization of the distillate.

It now is obtained largely as a by-product of the butyl alcohol industry. This alcohol is formed in the fermentation of carbohydrates such as corn starch, molasses, etc, by the action of the bacterium Clostridium acetobutylicum (Weizmann fermentation), and it is always one of the products formed in the process. It also is obtained by the catalytic oxidation of isopropyl alcohol, which is prepared from propylene resulting from the cracking of crude petroleum.

Description—Transparent, colorless, mobile, volatile, flammable liquid with a characteristic odor; specific gravity not more than 0.789; distills between 55.5 and 57°C; congeals about -95°C; aqueous solution neutral to litmus.

Solubility—Miscible with water, alcohol, ether, chloroform, or most volatile oils.

Uses—An antiseptic in concentrations above 80%. In combination with alcohol it is used as an antiseptic cleansing solution. It is employed as a menstruum in the preparation of oleoresins in place of ether. It is used as a solvent for dissolving fatty bodies, resins, pyroxylin, mercurials, etc, and also in the manufacture of many organic compounds such as chloroform, chlorobutanol, and ascorbic acid.

ALCOHOL

Ethanol; Spiritus Vini Rectificatus; S. V. R.; Spirit of Wine; Methylcarbinol

Ethyl alcohol [64-17-5]; contains 92.3 to 93.8%, by weight (94.9 to 96.0%, by volume), at 15.56°C (60°F) of C_2H_5OH (46.07).

Preparation—Has been made for centuries by fermentation of certain carbohydrates in the presence of zymase, an enzyme present in yeast cells. Usable carbohydrate-containing materials include molasses, sugar cane, fruit juices, corn, barley, wheat, potato, wood, and waste sulfite liquors. As yeast is capable of fermenting only D-glucose, D-fructose, D-mannose, and D-galactose, it is essential that more complex carbohydrates, such as starch, be converted to one or more of these simple sugars before they can be fermented. This is accomplished variously, commonly by enzyme- or acid-catalyzed hydrolysis.

The net reaction that occurs when a hexose, glucose for example, is fermented to alcohol may be represented as

$$C_6H_{12}O_6 \equiv 2\ C_2H_5OH + 2\ CO_2$$

but the mechanism of the process is very complex. The fermented liquid, containing about 15% alcohol, is distilled to obtain a distillate contain-

ing 94.9% C_2H_5OH, by volume. To produce absolute alcohol, various processes dehydrate the 95% product.

Hydration of ethylene, abundant supplies of which are available from natural and coke oven gases, from waste gases of the petroleum industry, and other sources may produce it also. In another synthesis acetylene is hydrated catalytically to acetaldehyde, which then is hydrogenated catalytically to ethyl alcohol.

Description—Transparent, colorless, mobile, volatile liquid; slight but characteristic odor; burning taste; boils at 78°C but volatilizes even at a low temperature, and is flammable; when pure, it is neutral toward all indicators; specific gravity at 15.56 (the US Government standard temperature for Alcohol) not above 0.816, indicating not less than 92.3% of C_2H_5OH by weight, or 94.9% by volume.

Solubility—Miscible with water, acetone, chloroform, ether, or many other organic solvents.

Incompatibilities—This and preparations containing a high percentage of alcohol will precipitate many inorganic salts from an aqueous solution. Acacia generally is recipitated from a hydroalcoholic medium when the alcohol content is greater than about 35%.

Strong oxidizing agents such as chlorine, nitric acid, permanganate, or chromate in acid solution react, in some cases violently, with it to produce oxidation products.

Alkali cause a darkening in color because of the small amount of aldehyde usually present in it.

Uses—In pharmacy principally for its solvent powers. It also is used as the starting point in the manufacture of many important compounds, like ether, chloroform, etc. It also is used as a fuel, chiefly in the denatured form.

It is a CNS depressant. Consequently, it occasionally has been administered intravenously for preoperative and postoperative sedation for patients in whom other measures are ineffective or contraindicated. The dose employed is 1 to 1.5 mL/kg. Its intravenous use is a specialized procedure and should be employed only by one experienced in the technique of such use.

It is used widely and abused by lay persons as a sedative. It has, however, no medically approved use for this purpose. Moreover, alcohol potentiates the CNS effects of numerous sedative and depressant drugs. Hence, patients taking certain prescription drugs or OTC medications should not use it.

Externally, it has a number of medical uses. It is a solvent for the toxicodendrol causing ivy poisoning and should be used to wash the skin thoroughly soon after contact. In a concentration of 25% it is employed for bathing the skin for the purpose of cooling and reducing fevers. In high concentrations it is a rubefacient and an ingredient of many liniments. In a concentration of 50% it is used to prevent sweating in astringent and anhidrotic lotions. It also is employed to cleanse and harden the skin and is helpful in preventing bedsores in bedridden patients. In a concentration of 60 to 90% it is germicidal. At optimum concentration (70% by weight) it is a good antiseptic for the skin (local antiinfective) and also for instruments. It also is used as a solvent to cleanse the skin splashed with phenol. High concentrations of it often are injected into nerves and ganglia for the relief of pain, accomplishing this by causing nerve degeneration.

DENATURED ALCOHOL

An act of Congress, June 7, 1906, authorizes the withdrawal of alcohol from bond without the payment of internal revenue tax, for the purpose of denaturation and use in the arts and industries. This is ethyl alcohol to which has been added such denaturing materials as to render the alcohol unfit for use as an intoxicating beverage. It is divided into two classes, namely, completely denatured alcohol and specially denatured alcohol, prepared in accordance with approved formulas prescribed in Federal Industrial Alcohol Regulations 3.

Completely Denatured Alcohol—This term applies to ethyl alcohol to which has been added materials (methyl isobutyl ketone, pyronate, gasoline, acetaldol, kerosene, etc) of such nature that the products may be sold and used within certain limitations without permit and bond.

Specially Denatured Alcohol—This alcohol is intended for use in a greater number of specified arts and industries than completely denatured alcohol, and the character of the denaturant or denaturants used is such that specially denatured alcohol may be sold, possessed, and used only by those persons or firms that hold basic permits and are covered by bond.

Uses—Approximately 50 specially denatured alcohol formulas containing combinations of more than 90 different denaturants are available to fill the needs of qualified users. Large amounts of specially denatured alcohols are used as raw materials in the production of acetaldehyde, synthetic rubber, vinegar, and ethyl chloride as well as in the manufacture of proprietary solvents and cleaning solutions. Ether and chloroform can be made from suitably denatured alcohols, and formulas for the manufacture of Iodine Tincture, Green Soap Tincture, and Rubbing Alcohol are set forth in the regulations.

Specially denatured alcohols also are used as solvents for surface coatings, plastics, inks, toilet preparations, and external pharmaceuticals. Large quantities are used in the processing of such food and drug products as pectin, vitamins, hormones, antibiotics, alkaloids, and blood products. Other uses include supplemental motor fuel, rocket and jet fuel, antifreeze solutions, refrigerants, and cutting oils. Few products are manufactured today that do not require the use of alcohol at some stage of production. Specially denatured alcohol may not be used in the manufacture of foods or internal medicines when any of the alcohol remains in the finished product.

DILUTED ALCOHOL

Diluted Ethanol

A mixture of alcohol and water containing 41.0 to 42.0%, by weight (48.4 to 49.5%, by volume), at 15.56°C, of C_2H_5OH (46.07).

Preparation

Alcohol	500 mL
Purified Water	500 mL

Measure the alcohol and the purified water separately at the same temperature, and mix. If the water and the alcohol and the resulting mixture are measured at 25°C, the volume of the mixture will be about 970 mL.

When equal volumes of alcohol and water are mixed together, a rise in temperature and a contraction of about 3% in volume take place. In small operations the contraction generally is disregarded; in larger operations it is very important. If 50 gal of official alcohol are mixed with 50 gal of water, the product will not be 100 gal of diluted alcohol, but only 96 1/4 gal, a contraction of 3 3/4 gal. US *Proof Spirit* differs from this and is stronger; it contains 50%, by volume, of absolute alcohol at 15.56°C (60°F). This corresponds to 42.5% by weight and has a specific gravity of 0.9341 at the same temperature. If spirits have a specific gravity lower than that of proof spirit (0.9341), they are said to be above proof; if greater, below proof.

It also may be prepared from the following:

Alcohol	408 g
Purified Water	500 g

Rules for Dilution—The following rules are applied when making an alcohol of any required lower percentage from an alcohol of any given higher percentage:

I. By Volume—Designate the volume percentage of the stronger alcohol by *V* and that of the weaker alcohol by *v*.

Rule—Mix *v* volumes of the stronger alcohol with purified water to make *V* volumes of product. Allow the mixture to stand until full contraction has taken place and until it has cooled, then make up the deficiency in the *V* volumes by adding more purified water.

Example—An alcohol of 30% by volume is to be made from an alcohol of 94.9% by volume.—Take 30 volumes of the 94.9% alcohol, and add enough purified water to produce 94.9 volumes at room temperature.

II. By Weight—Designate the weight-percentage of the stronger alcohol by *W* and that of the weaker alcohol by *w*.

Rule—Mix *w* parts by weight of the stronger alcohol with purified water to make *W* parts by weight of product.

Example—An alcohol of 50% by weight is to be made from an alcohol of 92.3% by weight.—Take 50 parts by weight of the 92.3% alcohol, and add enough purified water to produce 92.3 parts by weight.

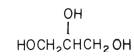
Description—As for Alcohol except its specific gravity is 0.935 to 0.937 at 15.56°C, indicating that the strength of C_2H_5OH corresponds to that given in the official definition.

Uses—A solvent in making tinctures, fluid-extracts, extracts, etc. Its properties already have been described fully in connection with the various preparations. Its value consists not only in its antiseptic properties, but also in its possessing the solvent powers of both water and alcohol. See Alcohol.

CHLOROFORM—page 1085.

GLYCERIN

1,2,3-Propanetriol; Glycerol



Glycerol [56-81-5] $C_3H_8O_3$ (92.09).

Chemically, it is the simplest trihydric alcohol. It is worthy of special note because the two terminal alcohol groups are primary, whereas the middle one is secondary. Thus this becomes the first polyhydric alcohol that can yield both an aldose (glyceraldehyde) and a ketose (dihydroxyacetone).

Preparation

1. By saponification of fats and oils in the manufacture of soap.
2. By hydrolysis of fats and oils through pressure and superheated steam.
3. By fermentation of beet sugar molasses in the presence of large amounts of sodium sulfite. Under these conditions a reaction takes place expressed as

$$\underset{\text{Glucose}}{C_6H_{12}O_6} \rightarrow \underset{\text{Glycerin}}{C_3H_5(OH)_3} + \underset{\text{Acetaldehyde}}{CH_3CHO} + CO_2$$

4. Glycerin is now prepared in large quantities from propylene, a petroleum product. This hydrocarbon is chlorinated at about 400°C to form allyl chloride, which is converted to allyl alcohol. Treatment of the unsaturated alcohol with hypochlorous acid (HOCl) yields the chlorohydrin derivative. Extraction of HCl with soda lime yields 2,3-epoxypropanol, which undergoes hydration to glycerin.

Description—Clear, colorless, syrupy liquid with a sweet taste and not more than a slight, characteristic odor, which is neither harsh nor disagreeable; when exposed to moist air it absorbs water and also such gases as H_2S and SO_2; solutions are neutral; specific gravity not below 1.249 (not less than 95% $C_3H_5(OH)_3$); boils at about 290°C under 1 atm, with decomposition, but can be distilled intact in a vacuum.

Solubility—Miscible with water, alcohol, or methanol; 1g in about 12 mL ethyl acetate or about 15 mL acetone; insoluble in chloroform, ether, or fixed and volatile oils.

Incompatibilities—An explosion may occur if it is triturated with strong oxidizing agents such as chromium trioxide, potassium chlorate, or potassium permanganate. In dilute solutions the reactions proceed at a slower rate, forming several oxidation products. Iron is an occasional contaminant of it and may be the cause of a darkening in color in mixtures containing phenols, salicylates, tannin, etc.

With boric acid or sodium borate, it forms a complex, generally spoken of as glyceroboric acid, which is a much stronger acid than boric acid.

Uses—One of the most valuable products known to pharmacy by virtue of its solvent property. It is useful as a humectant in keeping substances moist, owing to its hygroscopicity. Its agreeable taste and high viscosity adapt it for many purposes. Some modern ice collars and ice bags contain it and water hermetically sealed within vulcanized rubber bags. The latter are sterilized by dipping in a germicidal solution and are stored in the refrigerator until needed. It also has some therapeutic uses. In pure anhydrous form, it is used in the eye to reduce corneal edema and to facilitate ophthalmoscopic examination. It is used orally as an evacuant and, in 50 to 75% solution, as a systemic osmotic agent.

ISOPROPYL ALCOHOL—pages 221 and 1629.

METHYL ALCOHOL

Methanol; Wood Alcohol

CH_3OH

Methanol [67-56-1] CH_4O (32.04).

Caution—It is poisonous.

Preparation—By the catalytic reduction of carbon monoxide or carbon dioxide with hydrogen. A zinc oxide–chromium oxide catalyst is used commonly.

Description—Clear, colorless liquid; characteristic odor; flammable; specific gravity not more than 0.790; distills within a range of 63.5 and 65.7°C.

Solubility—Miscible with water, alcohol, ether, benzene, or most other organic solvents.

Uses—A solvent for non-ingested preparations. It is toxic. Ingestion may result in blindness; vapors also may cause toxic reactions.

METHYL ISOBUTYL KETONE

2-Pentanone, 4-methyl-, $(CH_3)_2CHCH_2COCH_3$ [108-10-1]; contains not less than 99% of $C_6H_{12}O$ (100.16).

Description—Transparent, colorless, mobile, volatile liquid; faint, ketonic and camphoraceous odor, distills between 114 and 117°C.

Solubility—Slightly soluble in water; miscible with alcohol, ether, or benzene.

Uses—A denaturant for rubbing alcohol and also a *solvent* for gums, resins, nitrocellulose, etc. It may be irritating to the eyes and mucous membranes, and, in high concentrations, narcotic.

MONOETHANOLAMINE

Ethanol, 2-amino-, Ethanolamine; Ethylolamine

$HOCH_2CH_2NH_2$ [141-43-5] C_2H_7NO (61.08).

Preparation—This alkanolamine is prepared conveniently by treating ethylene oxide with ammonia.

Description—Clear, colorless, moderately viscous liquid; distinctly ammoniacal odor; affected by light; specific gravity 1.013 to 1.016; distills between 167 and 173°C.

Solubility—Miscible in all proportions with water, acetone, alcohol, glycerin, or chloroform; immiscible with ether, solvent hexane, or fixed oils; dissolves many essential oils.

Uses—A *solvent* for fats, oils, and many other substances, it is a pharmaceutical necessity for Thimerosal Solution (see RPS-17 page 1173). It combines with fatty acids to form soaps that find application in various types of emulsions such as lotions, creams, etc.

PROPYLENE GLYCOL

$CH_3CH(OH)CH_2OH$

1,2-Propanediol [57-55-6 $C_3H_8O_2$] (76.10).

Preparation—Propylene is converted successively to its chlorohydrin (with HOCl), epoxide (with Na_2CO_3), and glycol (with water in presence of protons).

Description—Clear, colorless, viscous, and practically odorless liquid; slightly acrid taste; specific gravity 1.035 to 1.037; completely distills between 184 and 189°C; absorbs moisture from moist air.

Solubility—Miscible with water, alcohol, acetone, or chloroform; soluble in ether; dissolves many volatile oils; immiscible with fixed oils.

Uses—A solvent, preservative, and humectant. See Hydrophilic Ointment (page 1078).

TROLAMINE

Ethanol, 2,2′,2″-nitrilotris-, Triethanolamine

2,2′,2″-Nitrilotriethanol [102-71-6] $N(C_2H_4OH)_3$ (149.19); a mixture of alkanolamines consisting largely of triethanolamine, containing some diethanolamine [$NH(C_2H_4OH)_2$ = 105.14] and monoethanolamine [$NH_2C_2H_4OH$ = 61.08].

Preparation—Along with some mono- and diethanolamine, by the action of ammonia on ethylene oxide.

Description—Colorless to pale yellow, viscous, hygroscopic liquid; slight odor of ammonia; aqueous solution is very alkaline; melts about 21°C; specific gravity 1.120 to 1.128; a strong base and readily combines even with weak acids to form salts.

Solubility—Miscible with water or alcohol; soluble in chloroform; slightly soluble in ether or benzene.

Uses—In combination with a fatty acid, eg, oleic acid (see Benzyl Benzoate Lotion, 748), as an emulsifier. See Monoethanolamine.

WATER—page 1070.

OTHER PHARMACEUTICAL SOLVENTS

Alcohol, Dehydrated, BP, PhI [Dehydrated Ethanol; Absolute Alcohol]—Transparent, colorless, mobile, volatile liquid; characteristic odor; burning taste; specific gravity not more than 0.798 at 15.56°C; hygroscopic, flammable and boils about 78°C. Miscible with water, ether, or chloroform. *Uses:* A pharmaceutical solvent, antimicrobial preservative and penetration enhancer for topical preparations.

MISCELLANEOUS PHARMACEUTICAL NECESSITIES

The agents listed in this section comprise a heterogeneous group of substances with both pharmaceutical and industrial applications. Pharmaceutically, some of these agents are used as diluents, enteric coatings, excipients, and filtering agents and as ingredients in products considered in other chapters. Industrially, some of these agents are used in various chemical processes, in the synthesis of other chemicals, and in the manufacture of fertilizers, explosives, etc.

ACETIC ACID

Acetic acid; a solution containing 36 to 37%, by weight, of $C_2H_4O_2$ (60.05).

Preparation—By diluting with distilled water an acid of higher concentration, such as the 80% product, or more commonly glacial acetic acid, using 350 mL of the latter for the preparation of each 1000 mL of acetic acid.

Description—Clear, colorless liquid, having a strong characteristic odor and a sharply acid taste; specific gravity about 1.045; congeals about -14°C; acid to litmus.

Solubility—Miscible with water, alcohol, or glycerin.

Uses—In pharmacy as a solvent for making diluted acetic acid. It also is used as a starting point in the manufacture of many other organic compounds, eg, acetates, acetanilid, sulfonamides, etc. It is official primarily as a pharmaceutical necessity for the preparation of Aluminum Subacetate Solution.

DILUTED ACETIC ACID

Dilute Acetic Acid

A solution containing, in each 100 mL, 5.7 to 6.3 g of $C_2H_4O_2$.

Preparation

Acetic Acid	158 mL
Purified Water, a sufficient quantity to make	1000 mL

Mix the ingredients.

Note—This acid also may be prepared by diluting 58 mL of glacial acetic acid with sufficient purified water to make 1000 mL.

Description—Essentially the same properties, solubility, purity, and identification reactions as Acetic Acid, but its specific gravity is about 1.008, and it congeals about -2°C.

Uses—Bactericidal to many types of microorganisms and occasionally is used in 1% solution for surgical dressings of the skin. A 1% solution is spermicidal. It also is used in vaginal douches for the management of Trichomonas, Candida, and Haemophilus infections.

GLACIAL ACETIC ACID

Concentrated Acetic Acid; Crystallizable Acetic Acid; Ethanolic Acid; Vinegar Acid

CH_3COOH

Glacial acetic acid [64-19-7] $C_2H_4O_2$ (60.05).

Preparation—This acid is termed glacial because of its solid, glassy appearance when congealed. In one process it is produced by distillation of weaker acids to which has been added a water-entraining substance such as ethylene dichloride. In this method, referred to as azeotropic distillation, the ethylene dichloride distills out with the water before the acid distills over, thereby effecting concentration of the latter.

In another process the aqueous acid is mixed with triethanolamine and heated. The acid combines with the triethanolamine to form a triethanolamine acetate. The water is driven off first; then, at a higher temperature, the triethanolamine compound decomposes to yield this acid.

A greater part of the acid now available is made synthetically from acetylene. When acetylene is passed into this acid containing a metallic catalyst such as mercuric oxide, ethylidene diacetate is produced, which yields, upon heating, acetic anhydride and acetaldehyde. Hydration of the former and air oxidation of the latter yields this acid.

Description—Clear, colorless liquid; pungent, characteristic odor; when well diluted with water, it has an acid taste; boils about 118°C; congeals at a temperature not lower than 15.6°C, corresponding to a minimum of 99.4% of CH_3COOH; specific gravity about 1.05.

Solubility—Miscible with water, alcohol, acetone, ether, or glycerin; insoluble in carbon tetrachloride or chloroform.

Uses—A caustic and vesicant when applied externally and is often sold under various disguises as a corn solvent. It is an excellent solvent for fixed and volatile oils and many other organic compounds. It is used primarily as an acidifying agent.

ALUMINUM

Aluminum Al (26.98); the free metal in the form of finely divided powder. It may contain oleic acid or stearic acid as a lubricant. It contains not less than 95% Al and not more than 5% Acid-insoluble substances, including any added fatty acid.

Description—Very fine, free-flowing, silvery powder free from gritty or discolored particles.

Solubility—Insoluble in water or alcohol; soluble in hydrochloric and sulfuric acids or in solutions of fixed alkali hydroxides.

Uses—A protective. An ingredient in Aluminum Paste.

ALUMINUM MONOSTEARATE

Aluminum, dihydroxy (octadecanoato-*O*-)-,

Dihydroxy (stearato) aluminum [7047-84-9]; a compound of aluminum with a mixture of solid organic acids obtained from fats, and consists chiefly of variable proportions of aluminum monostearate and aluminum monopalmitate. It contains the equivalent of 14.5 to 16.5% of Al_2O_3 (101.96).

Preparation—By interaction of a hydroalcoholic solution of potassium stearate with an aqueous solution of potassium alum, the precipitate being purified to remove free stearic acid and some aluminum distearate simultaneously produced.

Description—Fine, white to yellowish white, bulky powder; faint, characteristic odor.

Solubility—Insoluble in water, alcohol, or ether.

Uses—A pharmaceutical necessity used in the preparation of Sterile Procaine Penicillin G with Aluminum Stearate Suspension.

STRONG AMMONIA SOLUTION

Stronger Ammonia Water; Stronger Ammonium Hydroxide Solution; Spirit of Hartshorn

Ammonia [1336-21-6]; a solution of NH_3 (17.03), containing 27.0 to 31.0% (w/w) of NH_3. Upon exposure to air it loses ammonia rapidly.

Caution—Use care in handling it because of the caustic nature of the Solution and the irritating properties of its vapor. Cool the container well before opening, and cover the closure with a cloth or similar material while opening. Do not taste it, and avoid inhalation of its vapor.

Preparation—Ammonia is obtained commercially chiefly by synthesis from its constituent elements, nitrogen and hydrogen, combined under high pressure and at high temperature in the presence of a catalyst.

Description—Colorless, transparent liquid; exceedingly pungent, characteristic odor; even when well diluted it is strongly alkaline to litmus; specific gravity about 0.90.

Solubility—Miscible with alcohol.

Uses—Only for chemical and pharmaceutical purposes. It is used primarily in making ammonia water by dilution and as a chemical reagent. It is too strong for internal administration. It is an ingredient in Aromatic Ammonia Spirit.

BISMUTH SUBNITRATE

Basic Bismuth Nitrate; Bismuth Oxynitrate; Spanish White; Bismuth Paint; Bismuthyl Nitrate

Bismuth hydroxide nitrate oxide [1304-85-4] $Bi_5O(OH)_9(NO_3)_4$ (1461.99); a basic salt that, dried at 105°C for 2 hr, yields upon ignition not less than 79% of Bi_2O_3 (465.96).

Preparation—A solution of bismuth nitrate is added to boiling water to produce the subnitrate by hydrolysis.

Description—White, slightly hygroscopic powder; suspension in distilled water is faintly acid to litmus (pH about 5).

Solubility—Practically insoluble in water or organic solvents; dissolves readily in an excess of hydrochloric or nitric acid.

Incompatibilities—Slowly hydrolyzed in *water* with liberation of nitric acid; thus, it possesses the incompatibilities of the acid. Reducing agents darken it with the production of metallic bismuth.

Uses—A pharmaceutical necessity in the preparation of milk of bismuth. It also is used as an astringent, adsorbent, and protective; however, its value as a protective is questionable. This agent, like other insoluble bismuth salts, is used topically in lotions and ointments.

BORIC ACID

Boric Acid (H_3BO_3); Boracic Acid; Orthoboric Acid

Boric acid [10043-35-3] H_3BO_3 (61.83).

Preparation—Lagoons of the volcanic districts of Tuscany formerly furnished the greater part of this acid and borax of commerce. Borax is now found native in California and some of the other western states; calcium and magnesium borates are found there also. It is produced from native borax or from the other borates by reacting with hydrochloric or sulfuric acid.

Description—Colorless scales of a somewhat pearly luster, or crystals, but more commonly a white powder slightly unctuous to the touch; odorless and stable in the air; volatilizes with steam.

Solubility—1 g in 18 mL water, 18 mL alcohol, 4 mL glycerin, 4 mL boiling water, or 6 mL boiling alcohol.

Uses—A buffer, and it is this use that is recognized officially. It is a very weak germicide (local anti-infective). Its nonirritating properties make its solutions suitable for application to such delicate structures as the cornea of the eye. Aqueous solutions are employed as an eyewash, mouth wash, and for irrigation of the bladder. A 2.2% solution is isotonic with lacrimal fluid. Solutions, even if they are made isotonic, will

hemolyze red blood cells. It also is employed as a dusting powder, when diluted with some inert material. It can be absorbed through irritated skin, eg, infants with diaper rash.

Although it is not absorbed significantly from intact skin, it is absorbed from damaged skin and fatal poisoning, particularly in infants, has occurred with topical application to burns, denuded areas, granulation tissue, and serous cavities. Serious poisoning can result from oral ingestion of as little as 5 g. Symptoms of poisoning are nausea, vomiting, abdominal pain, diarrhea, headache, and visual disturbance. Toxic alopecia has been reported from the chronic ingestion of a mouth wash containing it. The kidney may be injured, and death may result. Its use as a preservative in beverages and foods is prohibited by national and state legislation. There is always present the danger of confusing it with dextrose when compounding milk formulas for infants. Fatal accidents have occurred. For this reason boric acid in bulk is colored, so that it cannot be confused with dextrose.

It is used to prevent discoloration of physostigmine solutions.

CALCIUM HYDROXIDE

Slaked Lime; Calcium Hydrate

Calcium hydroxide [1305-62-0] $Ca(OH)_2$ (74.09).

Preparation—By reacting freshly prepared calcium oxide with water.

Description—White powder; alkaline, slightly bitter taste; absorbs carbon dioxide from the air, forming calcium carbonate; solutions exhibit a strong alkaline reaction.

Solubility—1 g in 630 mL water or 1300 mL boiling water; soluble in glycerin or syrup; insoluble in alcohol; the solubility in water is decreased by the presence of fixed alkali hydroxides.

Uses—In the preparation of Calcium Hydroxide Solution.

CALCIUM HYDROXIDE TOPICAL SOLUTION

Calcium Hydroxide Solution; Lime Water

A solution containing, in each 100 mL, not less than 140 mg of $Ca(OH)_2$ (74.09).

Note—The solubility of calcium hydroxide varies with the temperature at which the solution is stored, being about 170 mg/100 mL at 15°C and less at a higher temperature. The official concentration is based upon a temperature of 25°C.

Preparation

Calcium Hydroxide	3 g
Purified Water	1000 mL

Add the calcium hydroxide to 1000 mL of cool, purified water, and agitate the mixture vigorously and repeatedly during 1 hr. Allow the excess calcium hydroxide to settle. Dispense only the clear, supernatant liquid.

The undissolved portion of the mixture is not suitable for preparing additional quantities of the solution.

The object of keeping limewater over undissolved calcium hydroxide is to ensure a saturated solution.

Description—Clear, colorless liquid; alkaline taste; strong alkaline reaction; absorbs carbon dioxide from the air, a film of calcium carbonate forming on the surface of the liquid; when heated, it becomes turbid, owing to the separation of calcium hydroxide, which is less soluble in hot than in cold water.

Uses—It is too dilute to be effective as a gastric antacid. It is employed topically as a protective in various types of lotions. In some lotion formulations it is used with olive oil or oleic acid to form calcium oleate, which functions as an emulsifying agent. The USP classes it as an astringent.

CALCIUM STEARATE

Octadecanoic acid, calcium salt

Calcium stearate [1592-23-0]; a compound of calcium with a mixture of solid organic acids obtained from fats, and consists chiefly of variable proportions of stearic and palmitic acids [calcium stearate, $C_{36}H_{70}CaO_4$ = 607.03; calcium palmitate, $C_{32}H_{62}CaO_4$ = 550.92]; contains the equivalent of 9 to 10.5% of CaO (calcium oxide).

Preparation—By precipitation from interaction of solutions of calcium chloride and the sodium salts of the mixed fatty acids (stearic and palmitic).

Description—Fine, white to yellowish white, bulky powder; slight, characteristic odor; unctuous and free from grittiness.

Solubility—Insoluble in water, alcohol, or ether.

Uses—A lubricant in the manufacture of compressed tablets. It also is used as a conditioning agent in food and pharmaceutical products. Its virtually nontoxic nature and unctuous properties makes it ideal for these purposes.

CALCIUM SULFATE

Sulfuric acid, calcium salt (1:1); Gypsum; Terra Alba

Calcium sulfate (1:1) [7778-18-9] $CaSO_4$ (136.14); dihydrate [10101-41-4] (172.17).

Preparation—From natural sources or by precipitation from interaction of solutions of calcium chloride and a soluble sulfate.

Description—Fine, white to slightly yellow-white, odorless powder.

Solubility—Dissolves in diluted HCl; slightly soluble in water.

Uses—A diluent in the manufacture of compressed tablets. It is sufficiently inert that few undesirable reactions occur in tablets made with this substance. It also is used for making plaster casts and supports.

CARBON TETRACHLORIDE

Methane, tetrachloro-, Tetrachloromethane

Carbon tetrachloride [56-23-5] CCl_4 (153.82).

Preparation—One method consists of catalytic chlorination of carbon disulfide.

Description—Clear, colorless liquid; characteristic odor resembling that of chloroform; specific gravity 1.588 to 1.590; boils about 77°C.

Solubility—Soluble in about 2000 volumes water; miscible with alcohol, acetone, ether, chloroform, or benzene.

Uses—Officially recognized as a solvent. Formerly it was used as a cheap anthelmintic for the treatment of hookworm infections, but it causes severe injury to the liver if absorbed.

CARNAUBA WAX

Obtained from the leaves of *Copernicia cerifera* Mart (Fam *Palmae*).

Preparation—Consists chiefly of myricyl cerotate with smaller quantities of myricyl alcohol, ceryl alcohol, and cerotic acid. It is obtained by treating the leaf buds and leaves of *Copernicia cerifera,* the so-called Brazilian Wax Palm, with hot water.

Description—Light-brown to pale-yellow, moderately coarse powder; characteristic bland odor; free from rancidity; specific gravity about 0.99; melts about 84°C.

Solubility—Insoluble in water; freely soluble in warm benzene; soluble in warm chloroform or toluene; slightly soluble in boiling alcohol.

Uses—A pharmaceutical aid used as a polishing agent in the manufacture of coated tablets.

CELLULOSE ACETATE PHTHALATE

Cellulose, acetate, 1,2-benzenedicarboxylate

Cellulose acetate phthalate [9004-38-0]; a reaction product of the phthalic anhydride and a partial acetate ester of cellulose. When dried at 105°C for 2 hr, it contains 19 to 23.5% of acetyl (C_2H_3O) groups and 30 to 36.0% of phthalyl (*o*-carboxybenzoyl, $C_8H_5O_3$) groups.

Preparation—Cellulose is esterified by treatment with acetic and phthalic acid anhydrides.

Description—Free-flowing, white powder; may have a slight odor of acetic acid.

Solubility—Insoluble in water or alcohol; soluble in acetone or dioxane.

Uses—An enteric tablet-coating material. Coatings of this substance disintegrate because of the hydrolytic effect of the intestinal esterases, even when the intestinal contents are acid. *In vitro* studies indicate that cellulose acetate phthalate will withstand the action of artificial gastric juices for long periods of time but will disintegrate readily in artificial intestinal juices.

MICROCRYSTALLINE CELLULOSE

Cellulose [9004-34-6]; purified, partially depolymerized cellulose prepared by treating alpha cellulose, obtained as a pulp from fibrous plant material, with mineral acids.

Preparation—Cellulose is subjected to the hydrolytic action of 2.5 *N* HCl at the boiling temperature of about 105°C for 15 min, whereby amorphous cellulosic material is removed and aggregates of crystalline cellulose are formed. These are collected by filtration, washed with water and aqueous ammonia, and disintegrated into small fragments, often termed cellulose crystallites, by vigorous mechanical means such as a blender. US Pat 3,141,875.

Description—Fine, white, odorless, crystalline powder; consists of free-flowing, non-fibrous particles.

Solubility—Insoluble in water, dilute acids, or most organic solvents; slightly soluble in NaOH solution (1 in 20).

Uses A tablet diluent and disintegrant and dry binder. It can be compressed into self-binding tablets that disintegrate rapidly when placed in water.

Microcrystalline Cellulose and Sodium Carboxymethylcellulose, co-processed—A colloid-forming, attrited mixture of sub-micron microcrystalline cellulose and sodium carboxymethylcellulose. Description and Solubility: Tasteless, odorless, white to off-white, coarse to fine powder; pH (dispersion) 6 to 8; swells in water, producing, when dispersed, a white, opaque dispersion or gel. Insoluble in organic solvents or dilute acids. Uses: Pharmaceutical aid (suspending agent). Resultant viscosities vary depending upon the grade used and the type of CMC present.

POWDERED CELLULOSE—pages 1074 and 1278.

CHLOROFORM

Methane, trichloro-,

Trichloromethane [67-66-3] $CHCl_3$ (119.38); contains 99 to 99.5% $CHCl_3$, the remainder consisting of alcohol.

Caution—Care should be taken not to vaporize it in the presence of a flame, because of the production of harmful gases (hydrogen chloride and phosgene).

Preparation—Made by the reduction of carbon tetrachloride with water and iron and by the controlled chlorination of methane.

The pure compound readily decomposes on keeping, particularly if exposed to moisture and sunlight, resulting in formation of phosgene (carbonyl chloride $COCl_2$) and other products. The presence of a small amount of alcohol greatly retards or prevents this decomposition; hence, the requirement that it contain 0.5 to 1% of alcohol. The alcohol combines with any phosgene, forming ethyl carbonate, which is nontoxic.

Description—Clear, colorless, mobile liquid; characteristic, ethereal odor; burning, sweet taste; not flammable, but its heated vapors burn with a green flame; affected by light and moisture; specific gravity 1.474 to 1.478, indicating 99 to 99.5% of $CHCl_3$; boils about 61°C; not affected by acids but is decomposed by alkali hydroxide into alkali chloride and sodium formate.

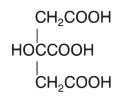
Solubility—Soluble in 210 volumes of water; miscible with alcohol, ether, benzene, solvent hexane, acetone, or fixed and volatile oils.

Uses—An obsolete inhalation anesthetic. Although it possesses advantages of non-flammability and great potency, it rarely is used because of the serious toxic effects it produces on the heart and liver. Internally, it has been used, in small doses, as a carminative. Externally, it is an *irritant* and when used in liniments it may produce blisters.

It is categorized as a pharmaceutical aid. It is used as a preservative during the aqueous percolation of vegetable drugs to prevent bacterial decomposition in the process of manufacture. In most instances it is evaporated before the product is finished. It is an excellent solvent for alkaloids and many other organic chemicals and is used in the manufacture of these products and in chemical analyses.

CITRIC ACID

1,2,3-Propanetricarboxylic acid, 2-hydroxy-,



Citric acid [77-92-9] $C_6H_8O_7$ (192.12); *monohydrate* [5949-29-1] (210.14).

Preparation—Found in many plants. It formerly was obtained solely from the juice of limes and lemons and from pineapple wastes. Since about 1925 the acid has been produced largely by fermentation of sucrose solution, including molasses, by fungi belonging to the *Aspergillus niger* group, theoretically according to the following reaction

$$\underset{\text{Sucrose}}{C_{12}H_{22}O_{11}} + \underset{\text{Oxygen}}{3O_2} \rightarrow \underset{\text{Citric Acid}}{2H_3C_6H_5O_7} + \underset{\text{Water}}{3H_2O}$$

but in practice there are deviations from this stoichiometric relationship.

Description—Colorless, translucent crystals, or a white, granular to fine crystalline powder; odorless; strongly acid taste; the hydrous form effloresces in moderately dry air but is slightly deliquescent in moist air; loses its water of crystallization at about 50°C; dilute aqueous solutions are subject to molding (fermentation), oxalic acid being one of the fermentation products.

Solubility—1 g in 0.5 mL water, 2 mL alcohol, or about 30 mL ether; freely soluble in methanol.

Uses—In the preparation of Anticoagulant Citrate Dextrose Solution, Anticoagulant Citrate Phosphate Dextrose Solution, Citric Acid Syrup, and effervescent salts. It also has been used to dissolve urinary bladder calculi and as a mild astringent.

COCOA BUTTER

Cacao Butter; Theobroma Oil; Oil of Theobroma

The fat obtained from the roasted seed of *Theobroma cacao* Linné (Fam *Sterculiaceae*).

Preparation—By grinding the kernels of the *chocolate bean* and expressing the oil in powerful, horizontal hydraulic presses. The yield is about 40%. It also has been prepared by dissolving the oil from the unroasted beans by the use of a volatile solvent.

Constituents—Chemically, it is a mixture of stearin, palmitin, olein, laurin, linolein, and traces of other glycerides.

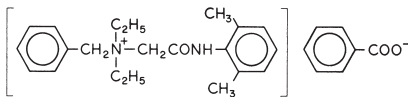
Description—Yellowish, white solid; faint, agreeable odor; bland (if obtained by extraction) or chocolate-like (if obtained by pressing) taste; usually brittle below 25°C; specific gravity 0.858 to 0.864 at 100°C/25°C; refractive index 1.454 to 1.458 at 40°C.

Solubility—Slightly soluble in alcohol; soluble in boiling dehydrated alcohol; freely soluble in ether or chloroform.

Uses—Valuable in pharmacy for making suppositories by virtue of its low fusing point and its property of becoming solid at a temperature just below the melting point. See Suppositories (page 883). In addition to this use, it is an excellent emollient application to the skin when inflamed; it also is used in various skin creams, especially the so-called skin foods. It also is used in massage.

DENATONIUM BENZOATE

Benzenemethanaminium *N*-2-(2,6-dimethylphenyl)amino-2-oxoethyl-*N,N*-diethyl-, benzoate;



Benzyldiethyl (2,6-xylylcarbamoyl)methylammonium benzoate [3734-33-6] $C_{28}H_{34}N_2O_3$ (446.59).

Preparation—2-(Diethylamino)-2',6'-xylidide is quaternized by reaction with benzyl chloride. The quaternary chloride then is treated with methanolic potassium hydroxide to form the quaternary base that, after filtering off the KCl, is reacted with benzoic acid. The starting xylidide may be prepared by condensing 2,6-xylidine with chloroacetyl chloride and condensing the resulting chloroacetoxylidide with diethylamine. US Pat. 3,080,327.

Description—White, odorless, crystalline powder; an intensely bitter taste; melts about 168°C.

Solubility—1 g in 20 mL water, 2.4 mL alcohol, 2.9 mL chloroform, or 5000 mL ether.

Uses—A denaturant for ethyl alcohol.

DEXTRIN

British Gum; Starch Gum; Leiocom

Dextrin [9004-53-9] $(C_6H_{10}O_5)_n$.

Preparation—By the incomplete hydrolysis of starch with dilute acid or by heating dry starch.

Description—White or yellow, amorphous powder (white: practically odorless; yellow: characteristic odor); dextrorotatory; $[\alpha]_D^{20}$ generally above 200°C; does not reduce Fehling's solution; gives a reddish color with iodine.

Solubility—Soluble in 3 parts of boiling water, forming a gummy solution; less soluble in cold water.

Uses—As a tablet diluent and an emulsifier in semi-solids.

DEXTROSE

Anhydrous Dextrose; Dextrose Monohydrate; Glucose; D(+)-Glucose; α-D(+)-Glucopyranose; Medicinal Glucose; Purified Glucose; Grape Sugar; Bread Sugar; Cerelose; Starch Sugar; Corn Sugar

D-Glucose monohydrate [5996-10-1] $C_6H_{12}O_6 \cdot H_2O$ (198.17); anhydrous [50-99-7] (180.16). A sugar usually obtained by the hydrolysis of starch.

Preparation—See Liquid Glucose (page 1086).

Description—Colorless crystals or a white, crystalline or granular powder; odorless; sweet taste; specific rotation (anhydrous) +52.5 to +53; anhydrous dextrose melts at 146°C; dextrose slowly reduces alkaline cupric tartrate TS in the cold and rapidly on heating, producing a red precipitate of cuprous oxide (difference from *sucrose*).

Solubility—1 g in 1 mL of water or 100 mL of alcohol; more soluble in boiling water or boiling alcohol.

Uses—See Dextrose Injection (page 1323). It also is used, instead of lactose as a supplement to milk for infant feeding.

DICHLORODIFLUOROMETHANE

Methane, dichlorodifluoro-, CCl_2F_2

Dichlorodifluoromethane [75-71-8] CCl_2F_2 (120.91).

Preparation—Carbon tetrachloride is reacted with antimony trifluoride in the presence of antimony pentafluoride.

Description—Clear, colorless gas; faint, ethereal odor; vapor pressure at 25°C about 4883 torr.

Uses—A propellant (No 12).

DICHLOROTETRAFLUOROETHANE

Ethane, 1,2-dichloro-1,1,2,2-tetrafluoro-, $CClF_2CClF_2$

1,2-Dichlorotetrafluoroethane [76-14-2] $C_2Cl_2F_4$ (170.92).

Preparation—By reacting 1,1,2-trichloro-1,2,2-trifluoroethane with antimony trifluorodichloride [SbF_3Cl_2], whereupon one of the 1-chlorine atoms is replaced by fluorine. The starting trichlorofluoroethane may be prepared from hexachloroethane by treatment with SbF_3Cl_2 (Henne AL: *Org Reactions II:* 65, 1944).

Description—Clear, colorless gas; faint, ethereal odor; vapor pressure at 25°C about 1620 torr; usually contains 6 to 10% of its isomer, $CFCl_2$-CF_3.

Uses—A propellant (No 114 and 114a).

EDETIC ACID

Glycine, *N,N′*-1,2-ethanediylbis[*N*-(carboxymethyl)], $(HOOCCH_2)_2NCH_2CH_2N(CH_2COOH)_2$

(Ethylenedinitrilo)tetraacetic acid [60-00-4] $C_{10}H_{16}N_2O_8$ (292.24).

Preparation—Ethylenediamine is condensed with sodium monochloroacetate with the aid of sodium carbonate. An aqueous solution of the reactants is heated to about 90°C for 10 hr, then cooled and acidified with HCl whereupon the acid precipitates. US Pat. 2,130,505.

Description—White, crystalline powder; melts with decomposition above 220°C.

Solubility—Very slightly soluble in water; soluble in solutions of alkali hydroxides.

Uses—A metal complexing agent. The acid, rather than any salt, is the form most potent in removing calcium from solution. It may be added to shed blood to prevent clotting. It also is used in pharmaceutical analysis and the removal or inactivation of unwanted ions in solution. Salts of the acid are known as edetates. See Edetate Calcium Disodium (page 1343) and Edetate Disodium (page 1343).

ETHYLCELLULOSE

Cellulose ethyl ether [9004-57-3]; an ethyl ether of cellulose containing 44 to 51% of ethoxy groups. The medium-type viscosity grade contains less than 46.5% ethoxy groups; the standard-type viscosity grade contains 46.5% or more ethoxy groups.

Preparation—By the same general procedure described on page ____ for Methylcellulose except that ethyl chloride or ethyl sulfate is employed as the alkylating agent. The 45 to 50% of ethoxy groups in the official ethylcellulose corresponds to from 2.25 to 2.61 ethoxy groups/$C_6H_{10}O_5$ unit, thus representing from 75 to 87% of the maximum theoretical ethoxylation, which is 3 ethoxy groups/$C_6H_{10}O_5$ unit.

Description—Free-flowing, white to light tan powder; forms films that have a refractive index of about 1.47; aqueous suspensions are neutral to litmus.

Solubility—The medium type is freely soluble in tetrahydrofuran, methyl acetate, chloroform, or mixtures of aromatic hydrocarbons with alcohol; the standard type is freely soluble in alcohol, methanol, toluene, chloroform, or ethyl acetate; both types are insoluble in water, glycerin, or propylene glycol.

Uses—A tablet binder and for film-coating tablets and drug particles.

GELATIN—page 1074.

LIQUID GLUCOSE

Glucose; Starch Syrup; Corn Syrup

A product obtained by the incomplete hydrolysis of starch. It consists chiefly of dextrose [D-(+)-glucose, $C_6H_{12}O_6$ = 180.16] dextrins, maltose, and water.

Preparation—Commercially by the action of very weak H_2SO_4 or HCl on starch.

One of the processes for its manufacture is as follows: The starch, usually from corn, is mixed with 5 times its weight of water containing less than 1% of HCl, the mixture is heated to about 45°C and then transferred to a suitable reaction vessel, into which steam is passed under pressure until the temperature reaches 120°C. The temperature is maintained at this point for about 1 hr or until tests show complete disappearance of starch. The mass is then heated to volatilize most of the hydrochloric acid, sodium carbonate or calcium carbonate is added to neutralize the remaining traces of acid, the liquid is filtered, then decolorized in charcoal or bone-black filters, as is done in sugar refining, and finally concentrated in vacuum to the desired consistency.

When made by the above process, it contains about 30 to 40% of dextrose mixed with about an equal proportion of dextrin, together with small amounts of other carbohydrates, notably maltose. By varying the conditions of hydrolysis, the relative proportions of the sugars also vary.

If the crystallizable dextrose is desired, the conversion temperature is higher, and the time of conversion longer. The term glucose, as customarily used in the chemical or pharmaceutical literature, usually refers to dextrose, the crystallizable product.

The name grape sugar sometimes is applied to the solid commercial form of dextrose because the principal sugar of the grape is dextrose, although the fruit has never been used as a source of the commercial supply.

Description—Colorless or yellowish, thick, syrupy liquid; odorless, or nearly so; sweet taste; differs from sucrose in that it readily reduces hot alkaline cupric tartrate TS, producing a red precipitate of cuprous oxide.

Solubility—Miscible with water; sparingly soluble in alcohol.

Uses—As an ingredient of Cocoa Syrup (page 1070), as a tablet binder and coating agent, and as a diluent in pilular extracts; it has replaced glycerin in many pharmaceutical preparations. It is sometimes given per rectum as a food in cases when feeding by stomach is impossible. It should not be used in the place of dextrose for intravenous injection.

HYDROCHLORIC ACID

Chlorhydric Acid; Muriatic Acid; Spirit of Salt

Hydrochloric acid [7647-01-0] HCl (36.46); contains 36.5 to 38.0%, by weight, of HCl.

Preparation—By the interaction of NaCl and H_2SO_4 or by combining chlorine with hydrogen. It is obtained as a by-product in the manufacture of sodium carbonate from NaCl by the Leblanc process in which common salt is decomposed with H_2SO_4. HCl is also a by-product in the electrolytic production of NaOH from NaCl.

Description—Colorless, fuming liquid; pungent odor; fumes and odor disappear when it is diluted with 2 volumes of water; strongly acid to litmus even when highly diluted; specific gravity about 1.18.

Solubility—Miscible with water or alcohol.

Uses—Officially classified as a pharmaceutical aid that is used as an acidifying agent. It is used in preparing Diluted Hydrochloric Acid.

HYPOPHOSPHOROUS ACID

Phosphoric acid

Hypophosphorous acid [6303-21-5] HPH_2O_2 (66.00); contains 30 to 32% by weight, of H_3PO_2.

Preparation—By reacting barium or calcium hypophosphite with sulfuric acid or by treating sodium hypophosphite with an ion-exchange resin.

Description—Colorless or slightly yellow, odorless liquid; solution is acid to litmus even when highly diluted; specific gravity about 1.13.

Solubility—Miscible with water or alcohol.

Incompatibilities—Oxidized on exposure to air and by nearly all oxidizing agents. Mercury, silver, and bismuth salts are reduced partially to the metallic state as evidenced by a darkening in color. Ferric compounds are changed to ferrous.

Uses—An antioxidant in pharmaceutical preparations.

ISOPROPYL MYRISTATE

Tetradecanoic acid, 1-methylethyl ester

$CH_3(CH_2)_{12}COOCH(CH_3)_2$

Isopropyl myristate [110-27-0] $C_{17}H_{34}O_2$ (270.45).

Preparation—By reacting myristoyl chloride with 2-propanol with the aid of a suitable dehydrochlorinating agent.

Description—Liquid of low viscosity; practically colorless and odorless; congeals about 5°C and decomposes at 208°C; withstands oxidation and does not become rancid readily.

Solubility—Soluble in alcohol, acetone, chloroform, ethyl acetate, toluene, mineral oil, castor oil, or cottonseed oil; practically insoluble in water, glycerin, or propylene glycol; dissolves many waxes, cholesterol, or lanolin.

Uses—Pharmaceutical aid used in cosmetics and topical medicinal preparations as an emollient, as a lubricant, and to enhance absorption through the skin.

KAOLIN—page 1313.

LACTIC ACID

Propanoic acid, 2-hydroxy-, 2-Hydroxypropionic Acid; Propanoloic Acid; Milk Acid

$CH_3CH(OH)COOH$

Lactic acid [50-21-5] $C_3H_6O_3$ (90.08); a mixture of lactic acid and lactic acid lactate ($C_6H_{10}O_5$) equivalent to a total of 85 to 90%, by weight, of $C_3H_6O_3$.

Discovered by Scheele in 1780, it is the acid formed in the souring of milk, hence the name *lactic,* from the Latin name for milk. It results from the decomposition of the lactose (milk sugar) in milk.

Preparation—A solution of glucose or of starch previously hydrolyzed with diluted sulfuric acid is inoculated, after the addition of suitable nitrogen compounds and mineral salts, with *Bacillus lactis*. Calcium carbonate is added to neutralize the lactic acid as soon as it is formed, otherwise the fermentation stops when the amount of acid exceeds 0.5%. When fermentation is complete, as indicated by failure of the liquid to give a test for glucose, the solution is filtered, concentrated, and allowed to stand. The calcium lactate that crystallizes is decomposed with dilute sulfuric acid and filtered with charcoal. The lactic acid in the filtrate is extracted with ethyl or isopropyl ether, the ether is distilled off, and the aqueous solution of the acid is concentrated under reduced pressure.

Description—Colorless or yellowish, nearly odorless, syrupy liquid; acid to litmus; absorbs water on exposure to moist air; when a dilute solution is concentrated to above 50%, lactic acid lactate begins to form; in the official acid the latter amounts to about 12 to 15%; specific gravity about 1.20; decomposes when distilled under normal pressure but may be distilled without decomposition under reduced pressure.

Solubility—Miscible with water, alcohol, or ether; insoluble in chloroform.

Uses—In the preparation of Sodium Lactate Injection (page 1341). It also is used in babies' milk formulas, as an acidulant in food preparations, and in 1 to 2% concentrations in some spermatocidal jellies. A 10% solution is used as a bactericidal agent on the skin of neonates. It is corrosive to tissues on prolonged contact. A 16.7% solution in flexible collodion is used to remove warts and small cutaneous tumors.

LACTOSE

D-Glucose, 4-*O*-β-D-galactopyranosyl-, Milk Sugar

Lactose [63-42-3] $C_{12}H_{22}O_{11}$ (342.30); monohydrate [10039-26-6] (360.31); a sugar obtained from milk.

Preparation—From skim milk, to which is added diluted HCl to precipitate the casein. After removal of the casein by filtration, the reaction of the whey is adjusted to a pH of about 6.2 by addition of lime, and heating coagulates the remaining albuminous matter; this is filtered out and the liquid set aside to crystallize. Animal charcoal is used to decolorize the solution in a manner similar to that used in purifying sucrose.

Another form of lactose, known as β-lactose, also is available on the market. It differs in that the D-glucose moiety is β instead of α. It is reported that this variety is sweeter and more soluble than ordinary lactose and for that reason is preferable in pharmaceutical manufacturing where lactose is used. Chemically, β-lactose does not appear to differ from ordinary α-lactose. It is manufactured in the same way as α-lactose up to the point of crystallization, then the solution is heated to a temperature above 93.5°C, the temperature at which the α form is converted to the β variety. The β form occurs only as an anhydrous sugar, whereas the α variety may be obtained either in the anhydrous form or as a monohydrate.

Description—White or creamy white, hard, crystalline masses or powder; odorless; faintly sweet taste; stable in air, but readily absorbs odors; pH (1 in 10 solution) 4 to 6.5; specific rotation +54.8 to +55.5.

Solubility—1 g in 5 mL water or 2.6 mL boiling water; very slightly soluble in alcohol; insoluble in chloroform or ether.

Uses—A diluent in tablet formulation. The amorphous and monohydrate forms are used in wet granulation processing of materials, whereas the spray-dried anhydrous type is usually used in direct compression formulations. It is generally an ingredient of the medium used in penicillin production. It is used extensively as an addition to milk for infant feeding.

MAGNESIUM CHLORIDE

Magnesium chloride hexahydrate [7791-18-6] $MgCl_2 \cdot 6H_2O$ (203.30); anhydrous [7786-30-3] (95.21).

Preparation—By treating magnesite or other suitable magnesium minerals with HCl.

Description—Colorless, odorless, deliquescent flakes or crystals, which lose water when heated to 100°C and lose HCl when heated to 110°C; pH (1 in 20 solution in carbon dioxide-free water) 4.5 to 7.

Solubility—Very soluble in water; freely soluble in alcohol.

Uses—Electrolyte replenisher; pharmaceutical necessity for hemodialysis and peritoneal dialysis fluids.

MAGNESIUM STEARATE

Octadecanoic acid, magnesium salt

Magnesium stearate [557-04-0]. A compound of magnesium with a mixture of solid organic acids obtained from fats, which consists chiefly of variable proportions of magnesium stearate and magnesium palmitate. It contains the equivalent of 6.8 to 8.0% MgO (40.30).

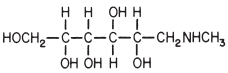
Description—Fine, white, bulky powder; faint, characteristic odor; unctuous, adheres readily to the skin and free from grittiness.

Solubility—Insoluble in water, alcohol, or ether.

Uses—A lubricant in the manufacture of compressed tablets.

MEGLUMINE

D-Glucitol, 1-deoxy-1-(methylamino)-,



1-Deoxy-1-(methylamino)-D-glucitol [6284-40-8] $C_7H_{17}NO_5$ (195.21).

Preparation—By treating glucose with hydrogen and methylamine under pressure and in the presence of Raney nickel.

Description—White to faintly yellowish white, odorless crystals or powder; melts about 130°C.

Solubility—Freely soluble in water; sparingly soluble in alcohol.

Uses—In forming salts of certain pharmaceuticals, surface-active agents and dyes. See Diatrizoate Meglumine Injections (page 1263), Iodipamide Meglumine Injection (page 1264) and Iothalamate Meglumine Injection (page 1266).

LIGHT MINERAL OIL

Light Liquid Petrolatum NF XII; Light Liquid Paraffin; Light White Mineral Oil

A mixture of liquid hydrocarbons obtained from petroleum. It may contain a suitable stabilizer.

Description—Colorless, transparent, oily liquid, free, or nearly free, from fluorescence; odorless and tasteless when cold, and develops not more than a faint odor of petroleum when heated; specific gravity 0.818 to 0.880; kinematic viscosity not more than 33.5 centistokes at 40°C.

Solubility—Insoluble in water or alcohol; miscible with most fixed oils, but not with castor oil; soluble in volatile oils.

Uses—Officially recognized as a vehicle. Once it was used widely as a vehicle for nose and throat medications; such uses are now considered dangerous because of the possibility of lipoid pneumonia. It sometimes is used to cleanse dry and inflamed skin areas and to facilitate removal of dermatological preparations from the skin. It should never be used for internal administration because of leakage. See Mineral Oil (page 1308).

NITRIC ACID

Nitric acid [7697-37-2] HNO_3 (63.01); contains about 70%, by weight, of HNO_3.

Preparation—May be prepared by treatment of sodium nitrate (Chile saltpeter) with sulfuric acid, but usually produced by catalytic oxidation of ammonia.

Description—Highly corrosive fuming liquid; characteristic, highly irritating odor; stains animal tissues yellow; boils about 120°C; specific gravity about 1.41.

Solubility—Miscible with water.

Uses—Acidifying agent

NITROGEN

Nitrogen [7727-37-9] N_2 (28.01); contains not less than 99%, by volume, of N_2.

Preparation—By the fractional distillation of liquefied air.

Uses—A diluent for medicinal gases. Pharmaceutically, is employed to replace air in the containers of substances that would be affected adversely by air oxidation. Examples include its use with fixed oils, certain vitamin preparations, and a variety of injectable products. It also is used as a propellant.

PHENOL

Carbolic Acid

C_6H_5OH

Phenol [108-95-2] C_6H_6O (94.11).

Preparation—For many years made only by distilling crude carbolic acid from coal tar and separating and purifying the distillate by repeated crystallizations; it now is prepared synthetically.

One process uses chlorobenzene as the starting point in the manufacture. The chlorobenzene is produced in a vapor phase reaction, with

benzene, HCl, and oxygen over a copper catalyst, followed by hydrolysis with steam to yield HCl and phenol (which is recovered).

Description—Colorless to light pink, interlaced, or separate, needle-shaped crystals, or a white or light pink, crystalline mass; characteristic odor; when undiluted, it whitens and cauterizes the skin and mucous membranes; when gently heated, phenol melts, forming a highly refractive liquid; liquefied by the addition of 10% of water; vapor is flammable; gradually darkens on exposure to light and air; specific gravity 1.07; boils at 182°C; congeals not lower than 39°C.

Solubility—1 g in 15 mL water; very soluble in alcohol, glycerin, chloroform, ether, or fixed and volatile oils; sparingly soluble in mineral oil.

Incompatibilities—Produces a liquid or soft mass when triturated with camphor, menthol, acetanilid, acetophenetidin, aminopyrine, antipyrine, ethyl aminobenzoate, methenamine, phenyl salicylate, resorcinol, terpin hydrate, thymol, and several other substances including some alkaloids. It also softens cocoa butter in suppository mixtures.

It is soluble in about 15 parts of water; stronger solutions may be obtained by using as much glycerin as phenol. Only the crystallized form is soluble in fixed oils and liquid petroleum, the liquefied form is not all soluble because of its content of water. Albumin and gelatin are precipitated by it. Collodion is coagulated by the precipitation of pyroxylin. Traces of iron in various chemicals such as alum, borax, etc, may produce a green color.

Uses—A caustic, disinfectant, topical anesthetic, and pharmaceutical necessity as a preservative for injections, etc. At one time widely used as a germicide and still the standard against which other antiseptics are compared, it has few legitimate uses in modern medicine. Nevertheless, it is still used in several proprietary antiseptic mouthwashes, hemorrhoidal preparations, and burn remedies. In full strength, a few drops of the liquefied form may be used to cauterize small wounds, dog bites, snake bites, etc. It commonly is employed as an antipruritic, in the form of phenolated calamine lotion (1%), phenol ointment (2%), or a simple aqueous solution (0.5 to 1%). It has been used for sclerosing hemorrhoids, but more effective and safer drugs are available. A 5% solution in glycerin is used in simple earache. Crude carbolic acid is an effective, economical agent for disinfecting excrement. It is of some therapeutic value as a fungicide, but more effective and less toxic agents are available. If accidentally spilled, it should be removed promptly from the skin by swabbing with alcohol.

Liquefied Phenol [Liquefied Carbolic Acid]—Phenol maintained in a liquid condition by the presence of 10.0% of water. It contains not less than 89.0%, by weight, of C_6H_6O. Note—When it is to be mixed with a fixed oil, mineral oil, or white petrolatum, use the crystalline Phenol, not Liquefied Phenol. Preparation: Melt phenol (a convenient quantity) by placing the unstoppered container in a steam bath and applying heat gradually. Transfer the liquid to a tared vessel, weigh, add 1 g of purified water for each 9 g of phenol, and mix thoroughly. Description: Colorless liquid, which may develop a red tint upon exposure to air and light; characteristic, somewhat aromatic odor; when undiluted it cauterizes and whitens the skin and mucous membranes; specific gravity about 1.065; when it is subjected to distillation, the boiling temperature does not rise above 182°C, which is the boiling temperature of phenol; partially solidifies at about 15°C. Solubility: Miscible with alcohol, ether, or glycerin; a mixture of liquefied phenol and an equal volume of glycerin is miscible with water. Uses: Its therapeutic uses are described above under Phenol. It is a pharmaceutical necessity for Phenolated Calamine Lotion (see RPS-18 page 762).

PHOSPHORIC ACID

Orthophosphoric Acid; Syrupy Phosphoric Acid; Concentrated Phosphoric Acid

Phosphoric acid [7664-38-2] H_3PO_4 (98.00); contains 85 to 88%, by weight, of H_3PO_4.

Preparation—Phosphorus is converted to phosphorus pentoxide P_2O_5 by exposing it to a current of warm air, then the P_2O_5 is treated with water to form phosphoric acid. The conversion of the phosphorus to the pentoxide takes place while the phosphorus, distilling from the phosphorus manufacturing operation, is in the vapor state.

Description—Colorless, odorless liquid of a syrupy consistency; specific gravity about 1.71.

Solubility—Miscible with water or alcohol, with the evolution of heat.

Uses—To make the diluted acid and as a weak acid in various pharmaceutical preparations. Industrially, it is used in dental cements and in beverages as an acidulant.

Diluted Phosphoric Acid [Dilute Phosphoric Acid]—Contains, in each 100 mL, 9.5 to 10.5 g of H_3PO_4 (98.00). Preparation: Mix phosphoric acid (69 mL) and purified water (qs) to make 1000 mL. Description and Solubility: Clear, colorless, odorless liquid; specific gravity about 1.057. Miscible with water or alcohol. Uses: A pharmaceutical necessity. It also has been employed in lead poisoning and in other conditions in which it is desired to administer large amounts of phosphate and at the same time produce a mild acidosis. It has been given in the dosage of 60 mL a day (5 mL/hr) under carefully controlled conditions.

POTASSIUM METAPHOSPHATE

Metaphosphoric acid (HPO_3), potassium salt

Potassium metaphosphate [7790-53-6] KPO_3 (118.07); a straight-chain polyphosphate, having a high degree of polymerization; contains the equivalent of 59 to 61% P_2O_5.

Preparation—By thermal dehydration of monopotassium phosphate (KH_2PO_4).

Description—White, odorless powder.

Solubility—Insoluble in water; soluble in dilute solutions of sodium salts.

Uses—Buffering agent

MONOBASIC POTASSIUM PHOSPHATE

For the full monograph, see page 1340.

Comments—A component of various buffer solutions. Medicinally, it has been used as a urinary acidifier.

PUMICE

Pumex

A substance of volcanic origin, consisting chiefly of complex silicates of aluminum, potassium, and sodium.

Description—Very light, hard, rough, porous, grayish masses or a gritty, grayish powder of several grades of fineness; odorless, tasteless, and stable in the air.

Three powders are available:

Pumice Flour or Superfine Pumice—Not less than 97% passes through a No 200 standard mesh sieve.

Fine Pumice—Not less than 95% passes through a Number 150 standard mesh sieve, and not more than 75% passes through a Number 200 standard mesh sieve.

Coarse Pumice—Not less than 95% passes through a No 60 standard mesh sieve, and not more than 5% passes through a No 200 standard mesh sieve.

Solubility—Insoluble in water and is not attacked by acids or alkali hydroxide solutions.

Uses—A filtering and distributing medium for pharmaceutical preparations. Because of its grittiness the powdered form is used in certain types of soaps and cleaning powders and also as a dental abrasive.

PYROXYLIN

Cellulose, nitrate; Soluble Guncotton

Pyroxylin [9004-70-0]; a product obtained by the action of a mixture of nitric and sulfuric acids on cotton, which consists chiefly of cellulose tetranitrate $(C_{12}H_{16}N_4O_{18})_n$.

Note—The commercially available form is moistened with about 30% of alcohol or other suitable solvent. The alcohol or solvent must be allowed to evaporate to yield the dried substance described in the *USP*.

Preparation—Shönbein, in 1846, found that nitric acid acts on cotton and produces a soluble compound. It subsequently was proved that this substance belongs to a series of closely related nitrates in which the nitric acid radical replaces the hydroxyl of the cellulose formula. Taking the double empirical formula for cellulose C12H20O10 and indicating replacement of four of the OH groups thus usually indicates this

$$\underset{\text{Cellulose}}{C_{12}H_{20}O_{10}} + 4HNO_3 \rightarrow \underset{\text{Cellulose Tetranitrate}}{C_{12}H_{16}O_6(NO_3)_4} + 4H_2O$$

The compound used in preparing collodion is a varying mixture of the di-, tri-, tetra-, and pentanitrates, but is mainly tetranitrate. The hexanitrate is the true explosive guncotton and is insoluble in ether, alcohol, acetone, or water.

Description—Light yellow, matted mass of filaments, resembling raw cotton in appearance but harsh to the touch; exceedingly flammable, burning, when unconfined, very rapidly and with a luminous flame; when kept in well-closed bottles and exposed to light, it is decomposed with the evolution of nitrous vapors, leaving a carbonaceous residue.

Solubility—Insoluble in water; dissolves slowly but completely in 25 parts of a mixture of 3 volumes of ether and 1 volume of alcohol; sol-

uble in acetone or glacial acetic acid and precipitated from these solutions by water.

Uses—A pharmaceutical necessity for collodion.

ROSIN

Resina; Colophony; Georgia Pine Rosin; Yellow Pine Rosin

A solid resin obtained from *Pinus palustris* Miller and from other species of *Pinus* Linné (Fam *Pinaceae*).

Constituents—American rosin contains sylvic acid [$C_{20}H_{30}O_2$], α-, β-, and γ-abietic acids [$C_{20}H_{30}O_2$], γ-pinic acid (from which α- and β-pinic acids are gradually formed), and resene. Some authorities also include pimaric acid [$C_{30}H_{20}O_2$] as a constituent. French rosin is called galipot.

Description—Sharply angular, translucent, amber-colored fragments, frequently covered with yellow dust; fracture brittle at ordinary temperatures, shiny and shallow conchoidal; odor and taste are slightly terebinthinate; easily fusible and burns with a dense, yellowish smoke, specific gravity 1.07 to 1.09

Solubility—Insoluble in water; soluble in alcohol, ether, benzene, glacial acetic acid, chloroform, carbon disulfide, dilute solutions of sodium hydroxide and potassium hydroxide, or some volatile and fixed oils.

Uses—A pharmaceutical necessity for Zinc-Eugenol Cement. Formerly, and to some extent still, used as a component of plasters, cerates, and ointments, to which it adds adhesive qualities.

PURIFIED SILICEOUS EARTH

Purified Kieselguhr; Purified Infusorial Earth; Diatomaceous Earth; Diatomite

A form of silica [SiO_2] [7631-86-9] consisting of the frustules and fragments of diatoms, purified by boiling with acid, washing, and calcining.

Occurrence and Preparation—Large deposits of this substance are found in Virginia, Maryland, Nevada, Oregon, and California, usually in the form of masses of rocks, hundreds of feet in thickness. Under the microscope it is seen to consist largely of the minute siliceous frustules of diatoms. It must be purified carefully in a manner similar to that directed for Talc (page 1091) and thoroughly calcined. The latter treatment destroys the bacteria that are present in large quantities in the native earth.

Description—Very fine, white, light-gray or pale-buff mixture of amorphous powder and lesser amounts of crystalline polymorphs, including quartz and cristobalite; gritty, readily absorbs moisture and retains about four times its weight of water without becoming fluid.

Solubility—Insoluble in water, acids, or dilute solutions of alkali hydroxides.

Uses—Introduced into the USP as a distributing and filtering medium for aromatic waters; also suitable for filtration of elixirs. Like talc, it does not absorb active constituents.

COLLOIDAL SILICON DIOXIDE

Silica [7631-86-9] SiO_2 (60.08); a submicroscopic fumed silica prepared by the vapor-phase hydrolysis of a silicon compound.

Description—Light, white, non-gritty powder of extremely fine particle size (about 15 nm).

Solubility—Insoluble in water or acids (except hydrofluoric); dissolved by hot solutions of alkali hydroxides.

Uses—A tablet moisture scavenger or glidant and as a suspending and thickening agent in non-solid preparations (coatings, semi-solids, liquids).

SODA LIME

A mixture of calcium hydroxide and sodium or potassium hydroxide or both.

It may contain an indicator that is inert toward anesthetic gases such as ether, cyclopropane, and nitrous oxide and that changes color when the soda lime no longer can absorb carbon dioxide.

Description—White or grayish white granules; if an indicator is added, it may have a color; absorbs carbon dioxide and water on exposure to air.

Uses—Neither a therapeutic nor a pharmaceutical agent. It is a reagent for the absorption of carbon dioxide in anesthesia machines, oxygen therapy, and metabolic tests. Because of the importance of the proper quality for these purposes it has been made official and standardized.

SODIUM BORATE

Sodium Tetraborate; Sodium Pyroborate; Sodium Biborate

Borax [1303-96-4] $Na_2B_4O_7 \cdot 10H_2O$ (381.37); anhydrous [1330-43-4] $Na_2B_4O_7$ (201.22).

Preparation—Found in immense quantities in California as a crystalline deposit. The earth, which is strongly impregnated with borax, is lixiviated (leached); the solution is evaporated and crystallized.

Calcium borate, or "cotton balls" also occurs in the borax deposits of California, and sodium borate is obtained from it by double decomposition with sodium carbonate.

Description—Colorless, transparent crystals, or a white, crystalline powder; odorless; the crystals often are coated with white powder because of efflorescence; solution alkaline to litmus and phenolphthalein; pH about 9.5.

Solubility—1 g in 16 mL water, 1 mL glycerin, or 1 mL boiling water; insoluble in alcohol.

Incompatibilities—Precipitates many metals as insoluble borates. In aqueous solution it is alkaline and precipitates aluminum salts as aluminum hydroxide, iron salts as a basic borate, and ferric hydroxide and zinc sulfate as zinc borate and a basic salt. Alkaloids are precipitated from solutions of their salts. Approximately equal weights of glycerin and boric acid react to produce a decidedly acid derivative generally called glyceroboric acid. Thus, the addition of glycerin to a mixture containing it overcomes incompatibilities arising from an alkaline reaction.

Uses—As a pharmaceutical necessity, it is used as an alkalizing agent and as a buffer for alkaline solutions. Its alkalizing properties provide the basis for its use in denture adhesives and its buffering action for its use in eyewash formulations.

SODIUM CARBONATE

Carbonic acid, disodium salt, monohydrate; Monohydrated Sodium Carbonate USP

Disodium carbonate monohydrate [5968-11-6] $Na_2CO_3 \cdot H_2O$ (124.00); anhydrous [497-19-8] (105.99).

Preparation—The initial process for its manufacture was devised by Leblanc, a French apothecary, in 1784, and consists of two steps: first, the conversion of common salt [NaCl] into sodium sulfate by heating it with sulfuric acid and, second, the decomposition of the sulfate by calcium carbonate (limestone) and charcoal (coal) at a high temperature to yield this salt and calcium sulfide. The carbonate then is leached out with water.

It currently is prepared by the electrolysis of sodium chloride, whereby sodium and chlorine are produced, the former reacting with water to produce sodium hydroxide and this solution treated with carbon dioxide to produce the salt. The process is used most extensively in localities where electric power is very cheap.

The monohydrated form is made by crystallizing a concentrated solution of this salt at a temperature above 35°C (95°F) and stirring the liquid so as to produce small crystals. It contains about 15% water of crystallization.

Soda ash is a term designating a commercial quality of the anhydrous salt. Its annual production is very large, and it has a wide variety of applications, among which are the manufacture of glass, soap, and sodium salts; it also is used for washing fabrics.

Description—Colorless crystals or a white, crystalline powder; stable in air under ordinary conditions; when exposed to dry air above 50°C it effloresces, and at 100°C it becomes anhydrous; decomposed by weak acids, forming the salt of the acid and liberating carbon dioxide; aqueous solution alkaline to indicators (pH about 11.5).

Solubility—1 g in 3 mL water or 1.8 mL boiling water; insoluble in alcohol.

Incompatibilities—Acids, acid salts, and acidic preparations cause its decomposition. Most metals are precipitated as carbonates, hydroxides, or basic salts. Alkaloids are precipitated from solutions of their salts.

Uses—Occasionally, for dermatitides topically as a lotion; it has been used as a mouthwash and a vaginal douche. It is used in the preparation of the sodium salts of many acids. The *USP* recognizes it as a pharmaceutical aid used as an alkalizing agent.

SODIUM HYDROXIDE

Caustic Soda; Soda Lye

Sodium hydroxide [1310-73-2] NaOH (40.00); includes not more than 3% Na_2CO_3 (105.99).

Caution—Exercise great care in handling it, as it rapidly destroys tissues—

Preparation—By treating sodium carbonate with milk of lime or by the electrolysis of a solution of sodium chloride as explained under Potassium Hydroxide (page 1287). It now is produced largely by the latter process. See also Sodium Carbonate, above.

Description—White, or nearly white, fused masses, small pellets, flakes, sticks, and other forms; hard and brittle and shows a crystalline fracture; exposed to the air, it rapidly absorbs carbon dioxide and moisture; melts at about 318°C; specific gravity 2.13; when dissolved in wa-

ter or alcohol or when its solution is treated with an acid, much heat is generated; aqueous solutions, even when highly diluted, are strongly alkaline.

Solubility—1 g in 1 mL water; freely soluble in alcohol or glycerin.

Incompatibilities—Exposed to air, it absorbs carbon dioxide and is converted to sodium carbonate. With fats and fatty acids it forms soluble soaps; with resins it forms insoluble soaps. See Potassium Hydroxide (page 1287).

Uses—Too alkaline to be of medicinal value but occasionally used in veterinary practice as a caustic. It is used extensively in pharmaceutical processes as an alkalizing agent and is generally preferred to potassium hydroxide because it is less deliquescent and less expensive; in addition, less of it is required, since 40 parts of it are equivalent to 56 parts of KOH. It is a pharmaceutical necessity in the preparation of Glycerin Suppositories.

SODIUM STEARATE

Octadecanoic acid, sodium salt

Sodium stearate [822-16-2] $C_{18}H_{35}NaO_2$ (306.47) consists chiefly of sodium stearate and sodium palmitate $C_{16}H_{31}NaO_2$ = 278.41).

Preparation—Stearic acid is reacted with an equimolar portion of NaOH.

Description—Fine, white powder, soapy to the touch; usually has a slight, tallow-like odor; affected by light; solutions are alkaline to phenolphthalein TS.

Solubility—Slowly soluble in cold water or cold alcohol; readily soluble in hot water or hot alcohol.

Uses—Officially, a pharmaceutical aid used as an emulsifying and stiffening agent. It is an ingredient of glycerin suppositories. In dermatological practice it has been used topically in sycosis and other skin diseases.

STARCH

Corn Starch; Wheat Starch; Potato Starch

Starch [9005-25-8] consists of the granules separated from the mature grain of corn *Zea mays* Linné (Fam *Gramineae*) or of wheat *Triticum aestivum* Linné (Fam *Gramineae*) or from tubers of the potato *Solanum tuberosum* Linné (Fam *Solanaceae*).

Preparation—In making starch from corn, the germ is separated mechanically, and the cells softened to permit escape of the starch granules. Permitting it to become sour and decomposed generally does this, stopping the fermentation before the starch is affected. On the small scale, making a stiff ball of dough and kneading it while a small stream of water trickles upon it may make it from wheat flour. It is carried off with the water, while the gluten remains as a soft, elastic mass; the latter may be purified and used for various purposes to which gluten is applicable. Commercially, its quality largely depends on the purity of the water used in its manufacture. It may be made from potatoes by first grating them, and then washing the soft mass upon a sieve, which separates the cellular substances and permits the starch granules to be carried through. It then must be washed thoroughly by decantation, and the quality of this starch also depends largely on the purity of the water that is used in washing it.

Description—Irregular, angular, white masses or fine powder; odorless; slight, characteristic taste. Corn starch: Polygonal, rounded, or spheroidal granules up to about 35 μm in diameter, which usually have a circular or several-rayed central cleft. Wheat starch: Simple lenticular granules 20 to 50 μm in diameter and spherical granules 5 to 10 μm in diameter, striations faintly marked and concentric. Potato starch: Simple granules, irregularly ovoid or spherical, 30 to 100 μm in diameter, and small spherical granules 10 to 35 μm in diameter; striations well marked and concentric.

Solubility—Insoluble in cold water or alcohol; when it is boiled with about 20 times its weight of hot water for a few minutes and then cooled, a translucent, whitish jelly results; aqueous suspension neutral to litmus.

Uses—Has absorbent and demulcent properties. It is used as a dusting powder and in various dermatological preparations; also as a pharmaceutical aid (filler, binder, and disintegrant). Note—Starches obtained from different botanical sources may not have identical properties with respect to their use for specific pharmaceutical purposes, eg, as a tablet-disintegrating agent. Therefore, types should not be interchanged unless performance equivalency has been ascertained.

Under the title Pregelatinized Starch, the NF recognizes starch that has been processed chemically or mechanically to rupture all or part of the granules in the presence of water and subsequently dried. Some types may be modified to render them compressible and flowable.

STORAX

Liquid Storax; Styrax; Sweet Gum; Prepared Storax

A balsam obtained from the trunk of *Liquidambar orientalis* Miller, known in commerce as Levant Storax, or of *Liquidambar styraciflua* Linné, known in commerce as American Storax (Fam *Hamamelidaceae*).

Constituents—The following occur in both varieties: styracin (cinnamyl cinnamate), styrol (phenylethylene, C_8H_8), α- and β-storesin (the cinnamic acid ester of an alcohol called storesinol), phenylpropyl cinnamate, free cinnamic acid, and vanillin. In addition to these, Levant storax contains ethyl cinnamate, benzyl cinnamate, free storesinol, isocinnamic acid, ethylvanillin, styrogenin, and styrocamphene. This variety yields from 0.5 to 1% of volatile oil; from this have been isolated styrocamphene, vanillin, the cinnamic acid esters of ethyl, phenylpropyl, benzyl, and cinnamyl alcohols, naphthalene, and styrol.

The American variety contains, in addition to the aforementioned substances common to both varieties, styaresin (the cinnamic acid ester of the alcohol styresinol, an isomer of storesinol) and styresinolic acid. It yields up to 7% of a dextrorotatory volatile oil, the composition of which has not been investigated completely; styrol and traces of vanillin have been isolated from it.

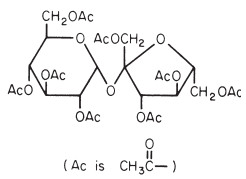
Description—Semi-liquid, grayish to grayish brown, sticky, opaque mass, depositing on standing a heavy dark brown layer (Levant storax); or a semisolid, sometimes a solid mass, softened by gently warming (American storax); transparent in thin layers; characteristic odor and taste; more dense than water.

Solubility—Insoluble in water, but soluble, usually incompletely, in an equal weight of warm alcohol; soluble in acetone, carbon disulfide, or ether, some insoluble residue usually remaining.

Uses—An expectorant but is used chiefly as a local remedy, especially in combination with benzoin; eg, it is an ingredient of Compound Benzoin Tincture. It may be used, like benzoin, to protect fatty substances from rancidity.

SUCROSE OCTAACETATE

α-D-Glucopyranoside, 1,3,4,6-tetra-*O*-acetyl-β-D-fructofuranosyl-, tetraacetate



Sucrose octaacetate [126-14-7] $C_{28}H_{38}O_{19}$ (678.60).

Preparation—Sucrose is subjected to exhaustive acetylation by reaction with acetic anhydride in the presence of a suitable condensing agent such as pyridine.

Description—White, practically odorless powder; intensely bitter taste; hygroscopic; melts not lower than 78°C.

Solubility—1 g in 1100 mL water, 11 mL alcohol, 0.3 mL acetone, or 0.6 mL benzene; very soluble in methanol or chloroform; soluble in ether.

Uses—A denaturant for alcohol.

SULFURATED POTASH

Thiosulfuric acid, dipotassium salt, mixed with potassium sulfide ($K_2(S_x)$); Liver of Sulfur

Dipotassium thiosulfate mixture with potassium sulfide (K_2S_x) [39365-88-3]; a mixture composed chiefly of potassium polysulfides and potassium thiosulfate. It contains not less than 12.8% S (sulfur) in combination as sulfide.

Preparation—By thoroughly mixing 1 part of sublimed sulfur with 2 parts of potassium carbonate and gradually heating the mixture in a covered iron crucible until the mass ceases to swell and is melted completely. It then is poured on a stone or glass slab and, when cold, is broken into pieces and preserved in tightly closed bottles. When the heat is regulated properly during its production, the reaction is represented approximately by

$$3K_2CO_3 + 8S \equiv 2K_2S_3 + K_2S_2O_3 + 3CO_2$$

As this product rapidly deteriorates on exposure to moisture, oxygen, and carbon dioxide, it is important that it be prepared recently to produce satisfactory preparations.

Description—Irregular pieces, liver-brown when freshly prepared, changing to a greenish yellow; decomposes upon exposure to air; an odor of hydrogen sulfide and a bitter, acrid, alkaline taste; even

weak acids cause the liberation of H_2S from sulfurated potash; 1 in 10 solution light brown in color and alkaline to litmus.

Solubility—1 g in about 2 mL water, usually leaving a slight residue; alcohol dissolves only the sulfides.

Uses—Extensively in dermatological practice, especially in the official White Lotion or *Lotio Alba* (page 1283). It is used as an opacifier.

The equation for the reaction of the potassium trisulfide in preparing the lotion is

$$ZnSO_4 + K_2S_3 \equiv ZnS + 2S + K_2SO_4$$

TALC

Talcum; Purified Talc; French Chalk; Soapstone; Steatite

A native, hydrous magnesium silicate, sometimes containing a small proportion of aluminum silicate.

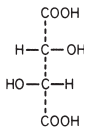
Occurrence and Preparation—The native form, called soapstone or French chalk, is found in various parts of the world. An excellent quality is obtained from deposits in North Carolina. Deposits of a high grade, conforming to the USP requirements, also are found in Manchuria. The native form usually is accompanied by variable amounts of mineral substances. These are separated from it by mechanical means, such as flotation or elutriation. It then is powdered finely, treated with boiling dilute HCl, washed well, and dried.

Description—Very fine, white, or grayish white crystalline powder, unctuous to the touch, adhering readily to the skin, and free from grittiness.

Uses—Officially, as a dusting powder and pharmaceutical aid; in both categories it has many specific uses. Its medicinal use as a dusting powder depends on its desiccant and lubricant effects. When perfumed, and sometimes medicated, it is used extensively for toilet purposes under the name talcum powder; for such use it should be in the form of an impalpable powder. When used as a filtration medium for clarifying liquids, a coarser powder is preferred to minimize passage through the pores of the filter paper; for this purpose it may be used for all classes of preparations with no danger of adsorption or retention of active principles. It is used as a lubricant in the manufacture of tablets and as a dusting powder when making handmade suppositories. Although it is used as a lubricant for putting on and removing rubber gloves, it should not be used on surgical gloves because even small amounts deposited in organs or healing wounds may cause granuloma formation.

TARTARIC ACID

Butanedioic acid, *R*-(*R**,*R**) 2,3-dihydroxy-,



L-(+)-Tartaric acid [87-69-4] $C_4H_6O_6$ (150.09).

Preparation—From argol, the crude cream of tartar (potassium bitartrate) deposited on the sides of wine casks during the fermentation of grapes, by conversion to calcium tartrate, which is hydrolyzed to tartaric acid and calcium sulfate.

Description—Large, colorless or translucent crystals, or a white granular to fine crystalline powder; odorless; acid taste; stable in the air; solutions acid to litmus; dextrorotatory.

Solubility—1 g in 0.8 mL water, 0.5 mL boiling water, 3 mL alcohol, or 250 mL ether; freely soluble in methanol.

Uses—Chiefly, as the acid ingredient of preparations in which it is neutralized by bicarbonate, as in effervescent salts and the free acid is completely absent or present only in small amounts in the finished product. It also is used as a buffering agent.

TITANIUM DIOXIDE—page 1293.

TRICHLOROMONOFLUOROMETHANE

Methane, trichlorofluoro-, $CFCl_3$

Trichlorofluoromethane [75-69-4] CCl_3F (137.37).

Preparation—Carbon tetrachloride is reacted with antimony trifluoride in the presence of a small quantity of antimony pentachloride. The reaction produces a mixture of CCl_3F and CCl_2F_2, which readily is separable by fractional distillation.

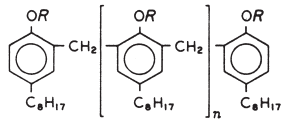
Description—Clear, colorless gas; faint, ethereal odor; vapor pressure at 25°C is about 796 torr; boils about 24°C.

Solubility—Practically insoluble in water; soluble in alcohol, ether, or other organic solvents.

Uses—A propellant.

TYLOXAPOL

Phenol, 4-(1,1,3,3-tetramethylbutyl)-, polymer with formaldehyde and oxirane



[*R* is $CH_2CH_2O(CH_2CH_2O)_mCH_2CH_2OH$; *m* is 6 to 8; *n* is not more than 5]

p-(1,1,3,3-Tetramethylbutyl)phenol polymer with ethylene oxide and formaldehyde [25301-02-4].

Preparation—*p*-(1,1,3,3-Tetramethylbutyl)phenol and formaldehyde are condensed by heating in the presence of an acidic catalyst, and the polymeric phenol thus obtained is reacted with ethylene oxide at elevated temperature under pressure in the presence of NaOH. US Pat. 2,454,541.

Description—Amber, viscous liquid; may show a slight turbidity; slight aromatic odor; specific gravity about 1.072; stable at sterilization temperature and in the presence of acids, bases, and salts; oxidized by metals; pH (5% aqueous solution) 4 to 7.

Solubility—Slowly but freely soluble in water; soluble in many organic solvents, including acetic acid, benzene, carbon tetrachloride, carbon disulfide, chloroform, or toluene.

Uses—A nonionic detergent that depresses both surface tension and interfacial tension. It also is used in contact-lens-cleaner formulations.

ISO-ALCOHOLIC ELIXIR

Iso-Elixir

Low-Alcoholic Elixir
High-Alcoholic Elixir of each a calculated volume
Mix the ingredients.

LOW-ALCOHOLIC ELIXIR

Compound Orange Spirit	10 mL
Alcohol	100 mL
Glycerin	200 mL
Sucrose	320 g
Purified Water qs	1000 mL

Alcohol Content—8 to 10%.

HIGH-ALCOHOLIC ELIXIR

Compound Orange Spirit	4 mL
Saccharin	3 g
Glycerin	200 mL
Alcohol, a sufficient quantity, to make	1000 mL

Alcohol Content—73 to 78%.

Uses Intended as a general vehicle for various medicaments that require solvents of different alcoholic strengths. When it is specified in a prescription, the proportion of its two ingredients to be used is that which will produce a solution of the required alcohol strength.

The alcoholic strength of the elixir to be used with a single liquid galenical in a prescription is approximately the same as that of the galenical. When galenicals of different alcoholic strengths are used in the same prescription, the elixir to be used is to be of such alcoholic strength as to secure the best solution possible. This generally will be found to be the average of the alcoholic strengths of the several ingredients.

For non-extractive substances, the lowest alcoholic strength of the elixir that will yield a perfect solution should be chosen.

UREA

For the full monograph, see page 1424.

Comments—A protein denaturant that promotes hydration of keratin and mild keratolysis in dry and hyperkeratotic skin. It is used in 2 to 20% concentrations in various dry-skin creams.

OTHER MISCELLANEOUS PHARMACEUTICAL NECESSITIES

Bucrylate [Propenoic acid, 2-cyano-, 2-methylpropyl ester; Isobutyl 2-cyanoacrylate [1069-55-2] $C_{80}H_{11}NO_2$ (153.18)]—Preparation: One method reacts isobutyl 2-chloroacrylate with sodium cyanide. Uses: Surgical aid (tissue adhesive).

Ceresin [Ozokerite; Earth Wax; Cerosin; Mineral Wax; Fossil Wax] A hard, white odorless solid resembling spermaceti when purified, occurring naturally in deposits in the Carpathian Mountains, especially in Galicia. It is a mixture of natural complex paraffin hydrocarbons. Melts between 61 and 78°C; specific gravity 0.91 to 0.92; stable toward oxidizing agents. Soluble in 30% alcohol, benzene, chloroform, petroleum, benzin, or hot oils. Uses: Substitute for beeswax; in dentistry, for impression waxes.

Ethylenediamine Hydrate BP, PhI [$H_2NCH_2CH_2NH_2 \bullet H_2O$]—Clear, colorless or slightly yellow liquid with an ammoniacal odor and characteristic alkaline taste; solidifies on cooling to a crystalline mass (mp 10°C); boils 118 to 119°C; specific gravity about 0.96; hygroscopic and absorbs CO_2 from the air; aqueous solutions alkaline to litmus. Miscible with water or alcohol; soluble in 130 parts of chloroform; slightly soluble in benzene or ether. Uses: In the manufacture of aminophylline and in the preparation of aminophylline injections. See Ethylenediamine (page 1059).

Ferric Oxide, Red—Contains not less than 90% Fe_2O_3. Heating native ferric oxide or hydroxide at a temperature that will yield a product of the desired color makes it. The temperature and time of heating, the presence and kind of other metals, and the particle size of the oxide govern the color. A dark-colored oxide is favored by prolonged heating at high temperature and the presence of manganese. A light-colored oxide is favored by the presence of aluminum and by finer particle size. Uses: Imparting color.

Ferric Oxide, Yellow—Contains not less than 97.5% Fe_2O_3. It is prepared by heating ferrous hydroxide or ferrous carbonate in air at a low temperature. Uses: Imparting color.

Honey NF [Mel; Clarified Honey; Strained Honey]—The saccharine secretion deposited in the honeycomb by the bee, *Apis mellifera* Linné (Fam *Apidae*). It must be free from foreign substances such as parts of insects, leaves, etc, but may contain pollen grains. Honey is one of the oldest of food and medicinal products. During the 16th and 17th Centuries it was recommended as a cure for almost everything. Constituents: invert sugar (62–83%), sucrose (0–8%), and dextrin (0.26–7%). Description: thick, syrupy liquid of a light yellowish to reddish brown color; translucent when fresh but frequently becomes opaque and granular through crystallization of dextrose; characteristic odor and a sweet, faintly acrid taste. Uses: A sweetening agent.

Polacrilin Potassium—Methacrylic acid polymer with divinylbenzene, potassium salt [39394-76-5]; Amberlite IRP-88. Prepared by polymerizing methacrylic acid with divinylbenzene, and the resulting resin is neutralized with KOH. Dry, buff-colored, odorless, tasteless, free-flowing powder; stable in light, air, and heat; insoluble in water. Uses: Pharmaceutical aid (tablet disintegrant ion exchange resin for controlled release liquids).

Poloxalene—Glycols, polymers, polyethylene-polypropylene [9003-11-6]. Polypropylene glycol is reacted with ethylene oxide. Uses: Pharmaceutical aid (surfactant).

Sodium Thioglycollate [Sodium Mercaptoacetate; $HSCH_2COONa$]—Hygroscopic crystals that discolor on exposure to air or iron. Freely soluble in water; slightly soluble in alcohol. *Uses:* Reducing agent in Fluid Thioglycollate Medium for sterility testing.

PART **6**

Pharmacokinetics and Pharmacodynamics

Paul Beringer, PharmD, BCPS
Associate Professor, Department of Pharmacy
USC School of Pharmacy
Los Angeles, CA

This page intentionally left blank.

CHAPTER 56

Diseases: Manifestations and Pathophysiology

Martin C Gregory, BM, BCh, DPhil
Michael B Strong, MD

This chapter provides a brief overview of certain basic information about some major diseases, the objective being to prepare students and practitioners of pharmacy for more effective service as drug information specialists and consultants on drug therapy.

We include symptoms and signs, pathophysiology, etiology and epidemiology of the diseases. Some discussion of relevant physiology, biochemistry, anatomy, and pathology serves to provide a better understanding of the diseases. Clinical features and means of diagnosis are discussed. Some conditions are discussed more extensively than others; many are not discussed at all. This uneven treatment is the result of variables such as state of knowledge, frequency of disease, applicability of drug therapy and space constraints. For additional information the reader should refer to textbooks of medicine or to textbooks of basic science disciplines for amplification of the introductory material provided here.

HEART DISEASE

Atherosclerosis

This is the single most important cause of mortality in the US because it is involved in the development of ischemic heart disease and cerebrovascular disease.

Normal Anatomy and Physiology—Arterial walls have three layers: the intima, media, and adventitia. The intima is composed of endothelial cells; the media of smooth muscle cells, and the adventitia of collagen, elastic fibers, fibroblasts, and some smooth muscle cells.

Arteries are not inert conduits, but metabolically complex structures that regulate their own caliber and perform many endothelial cell functions, including local inhibition of blood clotting and maintenance of cellular integrity. Throughout life, arteries withstand tremendous physical forces. Areas of particular stress, friction, and turbulence include bifurcations and openings of branch arteries.

Epidemiology—In all populations studied, early changes of atherosclerosis have been seen in young individuals who died of unrelated causes. Mortality and morbidity is more common in men than premenopausal women. After menopause, the differences decrease. The incidence of atherosclerotic disease also is different in various nationalities. The mortality from atherosclerotic disease in North Americans and Scots is twice that of Swedes. The incidence is low in Japanese and in native Africans. The incidence of atherosclerotic disease in immigrants to the US is similar to that of native Americans rather than to that of age-matched individuals who did not migrate. Primary relatives of individuals who become symptomatic from atherosclerotic disease before 50 years are likely to develop symptomatic atherosclerotic disease at an earlier age.

Etiology—Although the etiology is not known, clinical and epidemiological studies suggest that many factors contribute to the disease process. The two most important risk factors are advancing age and male sex. Other significant factors include diabetes mellitus, plasma cholesterol level, arterial blood pressure, cigarette smoking, and plasma homocysteine.

Other risk factors associated with a high incidence of atherosclerotic disease include diet, lack of physical activity, obesity, and heredity. The role of a competitive aggressive personality (Type A) is debated. These factors may not be independent of the others already listed.

Pathology—Atherosclerosis is a patchy thickening and hardening of arterial walls that is characterized in the early stages by streaks of cholesterol and other lipids ("fatty streaks") and later by atheromas. Atheromas consist of a fibrous cap that covers proliferating smooth muscle cells. When advanced they contain a necrotic core of lipids and proteins, the lesions initially involve the intima and progress to involve the media. Rupture of the fibrous cap precipitates thrombosis of the vessel.

Pathophysiology—The mechanism for the development is poorly understood, but increased stress associated with increased blood pressure and turbulent flow may foster development of lesions. The actual initiating event in the intima is unknown, but minute tears in this layer occur and may be important. Platelet aggregation and changes in endothelial permeability and fibrin deposition are important in the development of the atheroma.

These changes may induce smooth muscle proliferation in the intima with subsequent lipid accumulation. Another hypothesis advocates lipid deposition as the inciting and most important event. Later, fibrosis, calcification, hemorrhage, ulceration, and thrombosis develop causing eventual rupture or further lumen narrowing causing tissue blood flow to be critically reduced.

Blood lipids (cholesterol and triglycerides) are carried in combination with phospholipids and proteins, as lipoproteins. Acceleration of atherosclerosis correlates best with elevations of the LDL fraction (low-density lipoprotein), which is rich in cholesterol and poor in triglycerides. Elevation of the HDL fraction (high-density lipoprotein) protects against atherosclerosis.

Symptoms and Signs—Manifestations of atherosclerotic disease depend on the location and degree of impairment of blood flow to an organ. Atherosclerotic disease presents as sudden death (probably due to a ventricular arrhythmia), angina pectoris, myocardial infarction, cerebrovascular accident, dissecting aneurysm, thrombosis of a major vessel, ischemic renal disease, or peripheral vascular disease. Only those that are not discussed elsewhere will be discussed here. Peripheral vascular disease may cause intermittent claudication (pain in the legs precipitated by exercise and relieved by rest), leg pain at night, atrophy, and weakness of leg muscles, loss of pulses in the feet, neuropathy, extreme sensitivity to cold and eventually, dry gangrene. Atherosclerosis of the mesenteric vessels may cause abdominal pain that is precipitated by eating (abdominal angina), weight loss, diarrhea, and steatorrhea. Thrombosis of these vessels will cause bowel infarction. The diagnosis of atherosclerotic disease usually is based on symptoms and signs of reduced organ perfusion. Noninvasive studies and angiography are often helpful in defining the sites of vessel narrowing.

Coronary Artery Disease

This disease (CAD) also is referred to as ischemic heart disease (IHD). Inadequate oxygen supply for myocardial demand is

caused most commonly by coronary artery atherosclerotic disease. Other causes of decreased oxygen delivery to the myocardium include coronary embolism, coronary ostial stenosis in tertiary syphilis, and coronary artery spasm. Anemia, carboxyhemoglobinemia, and hypoxemia from lung disease can also reduce the oxygen-carrying capacity of the blood. Perfusion and O_2 delivery of the myocardium is decreased in hypotension. Myocardial oxygen demand is increased with exertion, myocardial hypertrophy, thyrotoxicosis, and beriberi. In the majority of cases of CAD, atherosclerosis is the underlying disorder.

Normal Anatomy—Arteries from the aorta nourish the myocardium. The right coronary artery supplies the right atrium, right ventricle, left atrium, posterior septum, AV node and, in over 50% of individuals, the SA node. The left coronary artery branches into two arteries. The circumflex supplies the anterolateral, lateral, posterolateral, inferior lateral, and inferior wall of the left ventricle, left atrium and, in about 45% of individuals, the SA node. The left anterior descending supplies the anterior, anterolateral, and apical left ventricular wall, the septum, and the right ventricle adjacent to the septum.

Normal Physiology—Under normal resting conditions, the myocardium extracts about 70% of the available oxygen from the coronary blood flow. This is in contrast to resting skeletal muscle, which extracts only 25% of available oxygen. Unlike skeletal muscle, the myocardium is capable of anaerobic metabolism for only a short time and cannot incur an oxygen debt. Increased myocardial oxygen demand must be met by increased coronary blood flow. The myocardium normally receives 5% of cardiac output. The normal heart can increase coronary blood flow by fivefold by a combination of coronary vasodilatation due to an autoregulatory process and increasing cardiac output. Blood flow to the myocardium occurs almost exclusively during diastole. Local tissue hypoxia results in potent vasodilatation and may increase coronary blood flow. Local tissue factors are more important than neuronal factors in regulating vasodilatation.

Systolic and diastolic wall tension, fraction of the cardiac cycle time spent in systole, and myocardial contractility determine myocardial oxygen consumption. The systolic wall tension (T) is determined by the systolic ventricular pressure (P), the radius (r) of the ventricular cavity and wal1 thickness (h) ($T = P\ r/2h$). The greater the cavity size and pressure, the greater the tension. Aortic diastolic pressure (afterload) partly determines the ventricular systolic pressure. The fraction of time spent in systole is determined by heart rate and ejection time. Oxygen demand of the myocardium depends on the amount of work the muscle must perform.

Pathology—Most atherosclerotic lesions occur in the proximal portion of the coronary arteries, because this is not a small vessel disease. Lesions in the left anterior descending artery are usually within 3 cm of the bifurcation of the left main coronary artery. Lesions in the right coronary artery usually occur within 6 to 8 cm of the ostium. A lesion that occludes less than 50% of the lumen of the vessel usually does not produce symptoms.

Pathophysiology—As the lumen of the vessels begins to narrow, blood flow decreases. Vessels distal to the obstruction dilate to maintain flow, presumably in response to hypoxia. When the obstruction reaches a critical size, the distal vessels become dilated permanently.

Ischemia causes changes in the biochemical, electrical, and mechanical properties of the heart. Myocardium normally oxidizes glucose and free fatty acids completely to carbon dioxide and water. In ischemia, lactate, pyruvate and other metabolic products accumulate in the myocardium. Ischemia also alters the electrical properties of the heart and decreases the membrane potential. Decreased conduction velocity and altered action potential duration result; thus arrhythmias may occur. Ischemia causes decreased contractility transiently, and necrosis causes irreversible loss of contractility. Ischemia may cause asymmetry and asynchrony of ventricular contraction.

The location of the lesion is important because this determines the size and location of the ischemia. The presence of collateral vessels may prevent the development of permanent injury. Unfortunately, the only known stimulus for collateral vessel formation is ischemia. A sudden decrease in lumen size, as with thrombosis or hemorrhage, is a catastrophic event as collateral vessels have not yet formed and therefore cannot provide an alternative source of oxygen. There is a group of people in whom coronary artery spasm plays a role in ischemic heart disease with or without fixed atherosclerotic lesions. How ischemia produces pain is unknown.

Angina pectoris is classified according to its frequency, severity, and precipitating event. Unstable angina describes a syndrome of attacks of recent onset or of increasing frequency, severity, or duration, or occurring with less exercise or at rest. Myocardial infarction and arrhythmias are more likely to develop during periods of unstable angina. Stable angina describes a clinical picture of little-varying attacks. Nocturnal angina occurs during sleep and may be associated with either dreams and rapid eye movement (REM) sleep or increased venous return in a patient with congestive failure. Prinzmetal angina is atypical angina. It occurs at rest, is associated with ventricular arrhythmias, and is thought to be due to coronary artery spasm.

Symptoms and Signs—CAD may present as ventricular arrhythmias or myocardial infarction, which will be discussed below. The other manifestation of CAD is angina pectoris, which is a clinical syndrome that results from transient myocardial ischemia but with no evidence of permanent damage.

The patient with angina pectoris usually describes the chest discomfort as heaviness, pressure, tightness, or squeezing. The patient often will not use the word “pain” or may ascribe his symptoms to indigestion. The substernal discomfort may radiate to the left arm, throat, jaw, shoulder, back, or abdomen. The discomfort typically is precipitated by exercise and also, but less often, by eating, emotional upset, exposure to cold, or cigarette smoking. Rest relieves it. The episodes usually last longer than 1 minute and not longer than 30 minutes.

Diagnosis—The diagnosis of angina is made from the history. The physical examination of these patients may be normal between attacks. Evidence of ischemia on EKG stress testing is inversion of T waves and depression of the S-T segment. Angiography and a therapeutic response to nitroglycerin may be helpful in establishing the diagnosis. Other causes of chest pain such as other forms of heart disease, gastrointestinal, and musculoskeletal disease must be considered in the differential diagnosis.

Ischemic heart disease is the leading cause of death among males over 35 years of age in the US and accounts for one-third of male deaths before age 65. The chief prognostic factors are the state of left ventricular function and the extent of the atherosclerotic disease.

Myocardial Infarction

Acute myocardial infarction (AMI) may be totally asymptomatic, can be fatal, or cause a variety of complications.

Pathology—The coronary arteries show thrombosis in approximately 90% of cases; thrombosis is central to the pathogenesis of AMI. Infarction is death of myocardial tissue. Most infarctions involve the endocardial layer. If the area of necrosis exceeds 3 cm in diameter, the infarct is likely to be transmural. Twenty-four hours after the infarction occurs, myocardial fibers show clumping, coagulation, and interstitial edema. By the 4th day the area is necrotic and shows fatty change and phagocytosis of fibers by neutrophils. Between the 4th and 10th days the area shows distinct fatty change, may contain hemorrhage, and is maximally soft. By the l0th day vascularized scar tissue begins to replace the infarct. The infarction heals completely by the 6th to 8th week.

Symptoms and Signs—Chest pain is usually the presenting complaint. It is described as severe, excruciating, deep, heavy, squeezing or crushing. No precipitating cause for the pain may be identified. The pain is similar to the pain of angina pectoris but is more severe, lasts longer, and is not relieved by rest or sublingual nitroglycerin. The pain may wax and wane. In 25% of patients, the substernal pain radiates to the arms; the pain also may radiate to the jaw, neck, abdomen, and back. Weakness, diaphoresis, nausea, vomiting, light-headedness, marked anxiety, and a sense of doom accompany the pain. The patient attempts in vain to find a comfortable position. In 15–20% of patients, AMI may be asymptomatic, particularly in diabetics. Elderly persons may complain of dyspnea rather than pain. Other presentations of MI include syncope, confusion, arrhythmias, and hypotension. Greater than 50% of the deaths following MI occur within the first 24 hours and are due to arrhythmias.

Physical examination typically discloses an anxious patient who is sweating and has cool extremities. Auscultation of the heart may reveal decreased heart sounds, S3, S4, or the murmur of mitral regurgitation. Temperature may be elevated to 38°.

Laboratory examination may reveal an increased white blood count to l5,000/mm^3. Enzymes released from damaged myocardial cells are used to diagnose AMI. The serum concentrations of these enzymes follow a characteristic pattern, with creatine (CK) and aspartate aminotransferase (AST) rising and falling quickly while lactic acid dehydrogenase (LDH) rises later and remains elevated longer. Measurement of serum troponin I has largely supplanted these enzyme measurements because of its greater sensitivity, specificity, and more rapid appearance after the infarction.

The EKG initially shows T-wave inversion and S-T segment elevation. When the infarct is transmural, Q-waves appear. Infarction also may cause decreased voltage in the precordial leads.

Complications—Arrhythmias are the most common cause of deaths in the early stages of AMI. Ventricular arrhythmias are the most ominous, and ventricular fibrillation is the most common fatal arrhythmia. Coronary care units that prevent or aggressively treat arrhythmias have decreased mortality from this complication.

Cardiac failure is now the primary cause of death in hospitalized patients with AMI. If greater than 40% of the myocardium is destroyed, the prognosis is poor.

Mitral regurgitation may occur as a result of rupture or dysfunction of the papillary muscles. This may decrease cardiac output and contribute to cardiac failure.

Thromboembolism contributes to the cause of death in some cases. Mural thrombosis may develop on the endocardium of the left ventricle. Emboli from this thrombus may cause strokes or a new AMI. Deep venous thrombosis may develop in the legs and embolize to the lungs (pulmonary embolism).

Rupture of the infarct may occur during the first week when the infarcted area is weakest. The blood pressure falls rapidly and the patient loses consciousness. The EKG may not change immediately. Cardiac tamponade occurs as the pericardium fills with blood. This complication is almost always fatal. The septum may rupture leading to left-to-right shunting. A pansystolic murmur is heard, and cardiac output decreases.

A weakening of the ventricular wall is called a ventricular aneurysm. When the remaining myocardium contracts, the aneurysm bulges. Because this portion of the wall has lost contractility, cardiac output decreases and congestive heart failure may develop. Systemic embolism may arise from a mural thrombus in the aneurysm. Arrhythmias are common with ventricular aneurysm and portend a poor prognosis.

Pericarditis may develop 2 to 3 days after the infarction. Pericardial pain is usually sharp, knife-like, and substernal. It may radiate to the neck and shoulders, is relieved by leaning forward, and is worsened by deep breathing. A pericardial friction rub may be heard. The pericarditis usually resolves with the healing of the infarct.

Heart Failure

This is defined as an inability of the heart, under normal filling conditions, to pump blood at a rate sufficient to meet the metabolic demands of the tissues. The inability to pump blood can be due to various abnormalities in the myocardium, coronary circulation, or heart valves. When the heart pumps blood at an insufficient rate, the kidneys retain salt and water and fluid accumulates in interstitial spaces. Thus, the term congestive heart failure usually is used. However, not all types of fluid overload or congestion are due to heart failure. Other causes of fluid overload include nephrotic syndrome, renal failure, liver disease, and starvation. Heart failure may develop acutely or chronically and may be mild to severe. Severe heart failure is synonymous with cardiogenic shock.

Normal Physiology—Function of the heart as a pump depends upon the number of functioning muscle fibers and their length at the onset of contraction. The end diastolic volume (EDV), which is referred to as preload, the cardiac impedance or afterload against which the blood is ejected, and the intrinsic myocardial activity or the contractile state determine fiber length. Heart rate and the stroke volume determine cardiac output (CO). The normal stroke volume (SV) is 70 ml, and the normal end systolic volume is 5 to 60 ml. SV is described by the equation: end diastolic volume minus end systolic volume.

The heart contracts in two phases. In the isovolumic phase, the length of the fiber remains constant while the pressure increases. When the left ventricular pressure reaches aortic diastolic pressure, the ejection phase occurs, during which contraction occurs as the fibers shorten. Heart rate determines filling time for the ventricles. With a normal heart, cardiac output remains stable between 50 to 180 beats/min. The afterload or the resistance against which the heart works influences cardiac output—the higher the resistance, the lower the CO.

Normally, the heart will pump out the blood that flows into it so that cardiac output is equal to venous return. Cardiac output can be increased within certain limits by autonomic stimulation, hypertrophy of the heart muscle, and an increase in blood volume. The heart has tremendous reserve capacity and can increase CO by increasing both heart rate and stroke volume.

The Frank-Starling relation describes the relationship between the stroke volume, diastolic volume or filling pressure and the length of the fibers at the end of diastole. The Frank-Starling relation also describes the ability of the heart to adapt to changing amounts of inflowing blood. Within physiological limits, the more the chamber is filled, the greater the quantity of blood that will be pumped. If the muscle fibers are stretched by volume, the muscle contracts with greater force, thereby increasing CO. Increased contractility results from sympathetic stimulation, and decreased contractility indicates a failing heart. This relationship shows that for a given end diastolic volume, the ventricles receive and eject a higher stroke volume when contractility is increased and a lower stroke volume when contractility is decreased.

Etiology—Processes that cause the heart to fail are those that increase the work of the heart, usually over many years, or that damage the myocardial fibers. As a result, cardiac output decreases. The most common cause of left ventricular failure is systemic hypertension. Stenotic (narrow) valves lead to heart failure sooner than incompetent (leaky) ones. Congenital defects may result in increased cardiac work. Cardiomyopathies, atherosclerotic coronary disease, and myocardial infarction damage muscle fibers and impair contractility. Tachyarrhythmias and atrial-ventricular dissociation reduce ventricular filling and ventricular arrhythmias decrease contractility. Pericarditis may impair ventricular filling. The most-common cause of right heart failure is left heart failure. Pulmonary embolism may precipitate acute right ventricular failure. Cor pulmonale is right heart failure due to pulmonary hypertension, which can occur as a complication of hypoxemia from lung disease.

Increased metabolic demands or decreased oxygen-carrying capacity of the blood may exceed cardiac reserve. Causes of high-output heart failure (see below) include hyperthyroidism, anemia, A-V fistulas, pregnancy, infections (particularly pulmonary infection), and beriberi.

Pathophysiology—The majority of cases of congestive heart failure (CHF) are due to low-output failure as occurs in hypertension, atherosclerotic heart disease, or valvular disease. Less commonly, the cardiac output is greater than normal, high-output heart failure. This is due to the metabolic demands of the tissue being increased greatly or the oxygen-carrying capacity of the blood being decreased greatly. Hyperthyroidism and pregnancy are causes of increased metabolic demands of the tissues. Anemia, arteriovenous fistulas, and hypoxemia are examples of decreased oxygen delivery.

Compensatory Mechanisms of Low Cardiac Output—When cardiac output falls, reflexes occur immediately. The baroreceptors sense the decreased arterial pressure and increase sympathetic tone while decreasing parasympathetic tone. This increases the force of contraction of the heart, increases heart rate, raises mean systemic arterial pressure, and increases venous return. These reflexes are maximal at 30 sec after a drop in arterial pressure.

Redistribution of blood flow occurs resulting in maintenance of blood flow to the myocardium and brain. Blood flows to the skin and skeletal muscle are decreased greatly by norepinephrine-induced vasoconstriction. Blood flow also is decreased to the kidneys in CHF.

Decreased cardiac output reduces the glomerular filtration rate because of both decreased renal blood flow and sympathetic vasoconstriction of afferent renal arterioles. Blood flow within the kidneys is redistributed by the vasoconstriction to the medulla at the expense of the cortex. Renin production by the juxtaglomerular apparatus is increased in response to decreased blood flow. Renin cleaves angiotensinogen to angiotensin I which is converted by angiotensin converting enzyme (ACE) to angiotensin II. Angiotensin II is a potent peripheral vessel constrictor and stimulator of aldosterone secretion by the adrenal cortex. Aldosterone promotes the retention of sodium and water by the distal convoluted renal tubule, causes expansion of the blood volume and accumulation of fluid in interstitial spaces. Aldosterone may have other harmful effects because blockade of its effects prolongs survival in heart failure. Serum sodium remains normal or is decreased, although total body sodium is increased.

Increased blood volume and increased systemic blood pressure increase venous return to the heart. Eventually, the heart can no longer keep pace with the increased venous return. In mild heart failure, the increased fluid volume helps to increase cardiac output by applying some stretch to the myocardial fibers. In severe heart failure, the amount of fluid overload becomes so great that the fibers are stretched beyond the limits of efficient contraction and the fibers descend to a lower Frank-Starling curve. A greater end-diastolic pressure is necessary to maintain CO on this lower curve. The increase in left ventricular end diastolic pressure (LVEDP) is transmitted as increased hydrostatic pressure to the pulmonary veins, capillaries, and arteries. Eventually, the increased pressure in the pulmonary arteries causes the right ventricle to fail. Increased right ventricular end diastolic pressure translates into increased hydrostatic pressure in the systemic veins and capillaries.

Edema Formation—Most of the fluid accumulation in interstitial spaces results from increases in capillary hydrostatic pressures. The colloidal osmotic (oncotic) pressure of the blood that holds fluid in the

vascular compartment is about 25 to 30 mmHg. When the hydrostatic pressure in the capillaries exceeds 25 mmHg, fluid is pushed into the interstitial spaces. CHF involves fluid retention in both the intra- and extravascular space. Fluid retention also results in distention of the venous reservoirs in the liver and spleen. When LVEDP exceeds 25 mmHg, the pressure transmitted to the pulmonary capillaries causes pulmonary edema. Oxygen does not diffuse efficiently in alveoli filled with edema fluid so hypoxemia results.

Symptoms and Signs—Patients with left ventricular failure most commonly complain of a sensation of shortness of breath (dyspnea). Initially, the dyspnea is present only on exertion (DOE), but the amount of activity necessary to precipitate dyspnea progressively lessens until the patient is dyspneic at rest. Orthopnea is the sensation of breathlessness that occurs in the recumbent position and may be relieved by elevating the head on several pillows or by sitting. Paroxysmal nocturnal dyspnea (PND) is severe dyspnea occurring at night that awakens the patient with a sensation of smothering. PND usually is accompanied by coughing and/or wheezing. The patient may produce frothy pink sputum.

Patients with left ventricular failure also may experience fatigue, weakness, and alterations in mental status such as confusion, difficulty in concentrating, impaired memory, headache, insomnia, and anxiety.

Physical examination of the patient with left heart failure reveals a person who may have lost considerable body mass. The patient may be unable to lie flat during the examination. The pulse may be weak, although the blood pressure remains normal until very late in the course. The extremities will be pale and cool. Cyanosis of the lips and nailbeds may be present. Examination of the heart reveals tachycardia and S3 gallop. Moist crepitant inspiratory crackles over the lung bases in moderately severe CHF and over the entire lung fields in pulmonary edema are heard. Chest radiograph shows the enlarged heart and signs of pulmonary venous congestion.

Patients with right ventricular failure complain of weight gain and the accumulation of fluid. A 10% gain in body weight may occur before pitting edema occurs. In ambulatory patients the edema is symmetrical in the ankles and legs. Since gravity influences the distribution of the edema, the buttocks and sacrum may be edematous in bedfast patients. Anasarca is massive body fluid overload including generalized edema, ascites, and pleural effusions. As fluids accumulate in the pleural cavity, the patient may develop dyspnea. Patients experience an increased girth as fluid accumulates in the peritoneal cavity. The liver may become enlarged and tender, and there may be right upper quadrant pain. Anorexia, nausea, and abdominal fullness occur. The patient becomes jaundiced as impairment of liver function becomes severe.

Physical examination of the patient with severe right-sided heart failure will reveal pleural effusions, ascites, jugular venous distention, hepatomegaly, splenomegaly, and pitting edema. Urine volume will be decreased, and prerenal azotemia may be present.

Valvular Heart Disease

This disease occurs when the heart valves become damaged and no longer will open or close properly. If fibrous scar tissue forms or calcium deposits on the valve, the valve becomes stenotic and no longer opens easily. If the valve leaflets shrink or do not oppose each other properly when closing, due to scarring or are destroyed by infection, the valve no longer is competent and blood flows in a retrograde fashion. A single valve may be both stenotic and incompetent. More than one valve may be involved. The consequences of valvular disease include congestive heart failure, arrhythmias, and systemic emboli. The hallmark of valvular disease is a murmur—a noise representing turbulence of blood flow across the valve.

Normal Anatomy and Physiology—The valves between the atria and the ventricles, tricuspid on the right side and mitral on the left side, are large and normally offer little resistance to flow. The semilunar valves, the aortic and pulmonic, are smaller. The atrioventricular valves are supported by chordae tendineae, but the semilunar valves are not. All valves open and close passively in response to pressure gradients.

The opening and closing of the valves cause the heart sounds. The first sound ("lub") occurs at the beginning of systole and represents the closure of the mitral and tricuspid valves. The second sound ("dup") is heard at the beginning of diastole and signals closure of the aortic and pulmonic valves. In normal individuals, the second sound may be split because the aortic and pulmonic valves do not close simultaneously. The first sound is loud when the mitral valve leaflets are far apart at the onset of ventricular systole. This occurs in mitral stenosis and tachycardia of any cause. A loud second sound indicates an increased pressure, as in systemic and pulmonary hypertension. A third heart sound (S3) is caused by blood flowing into the ventricles early in diastole, particularly into the dilated ventricles of CHF. A fourth heart sound (S4) also emanates from the ventricles late in diastole, but is caused by forceful atrial contraction thrusting blood into ventricles whose compliance is decreased, as in hypertensive heart disease.

Etiology—Formerly most valvular lesions followed rheumatic fever. Now the causes are more varied and include congenital lesions such as a bicuspid aortic valve, which may become significantly stenotic only in adult life as it calcifies. Another congenital condition is mitral valve prolapse, in which the mitral valve is redundant and billows into the left atrium during systole.

A number of systemic diseases are associated with valvular lesions: seronegative spondylitides, polycystic kidney disease, and Marfan's syndrome are examples. Acute bacterial endocarditis can attack previously normal valves and destroy them rapidly. The aortic and tricuspid valves are particularly vulnerable, especially in intravenous drug users. Subacute bacterial endocarditis also damages valves, but it usually alights on previously abnormal valves.

An increasingly common cause of mitral and tricuspid insufficiency is dilatation of the valve ring from chronic fluid overload as in CHF and end-stage renal disease.

Pathology—In acute rheumatic fever, the valve leaflets become swollen and thickened and small bead-like nodules develop along the valve closure lines and on the chordae. These nodules are composed of fibrin, platelets, and white blood cells. The inflammation may subside with the acute attack or develop into a subacute or chronic process. The inflammation leads to erosion of the endothelial surface and deposition of collagen by fibroblasts. The fibrous scarring during organization leaves a permanently thickened, distorted, rigid valve. Contraction of the scar results in shortening of the leaflets and distortion of the architecture of the valve. The edges of the deformed valve fail to fit together during closure causing valvular incompetence. The chordae also may be involved in scarring and shortening. Fibrous adhesions may occur across the cusp edges. Irregular fibrous thickening and scarring also are associated with calcification. Adhesions and calcification increase the rigidity of the valve and cause stenosis. Stenosis and the uneven surface are associated with increased turbulence of flow across the valve.

MITRAL STENOSIS (MS)

Pathophysiology—The normal area of the mitral valve is 4 to 6 cm^2 in adults. Symptoms of MS occur when the area is reduced to 1.5 cm^2. If the valve is stenotic, greater pressures are required to pump the blood from the left atrium to the left ventricle. Normally mean pressure in the left atrium is 12 mmHg. A valve orifice less than 1 cm^2 requires a pressure of 25 mmHg in the left atrium to pump the blood into the left ventricle. The elevated atrial pressure is transmitted back into the pulmonary veins, capillaries, and arteries. Pulmonary arteries develop medial hypertrophy and intimal thickening, which leads to high resistance and pulmonary hypertension. Alveolar fibrosis may occur also. When the pressure in the pulmonary capillaries exceeds the osmotic pressure of blood, pulmonary edema develops although the patient does not have left ventricular failure and left ventricular end diastolic pressure is normal. Eventually, the right heart fails.

Left ventricular output may be normal or decreased. The flow across the valve depends on heart rate as well as size of the opening. Increasing heart rate decreases the time available for flow across the mitral valve.

Symptoms and Signs—Two decades usually elapse between the initial attack of rheumatic fever and the development of the symptoms and signs of MS. Most patients become symptomatic during the fourth decade of life. Once the symptoms occur, the prognosis is poor with death occurring in 2 to 5 years unless the valve is replaced. Symptoms begin with dyspnea and cough during extreme exertion, but over the years the amount of exercise necessary to produce symptoms decreases until dyspnea occurs at rest. Orthopnea and paroxysmal nocturnal dyspnea may occur. With longstanding MS, atrial arrhythmias are common.

Extensive fibrosis of the alveolar walls and pulmonary capillary thickening lead to decreased vital capacity, total lung capacity, maximum breathing capacity and oxygen uptake. V/Q mismatching occurs. Decreased compliance of the lung increases the work of breathing and increases the sensation of breathlessness. Hemoptysis results from rupture of small vessels in the bronchioles.

Patients with MS, particularly those with atrial fibrillation, are likely to embolize thrombi from the left atrium to the brain, kidneys, spleen, or extremities.

The physical examination of patients with MS often discloses cyanosis of the lips and nails, and signs of right heart failure. The first heart sound is accentuated. The opening snap of the mitral valve may be heard. A low-pitched rumbling diastolic murmur is characteristic of mi-

tral stenosis. Chest radiograph shows an enlarged left atrium, pulmonary arteries, and right ventricle, as well as markings of increased pulmonary venous pressure. EKG shows signs of left atrial enlargement and may disclose an atrial arrhythmia. Echocardiogram, the most useful noninvasive test shows inadequate valve separation, thickened leaflets, and left atrial enlargement.

MITRAL INSUFFICIENCY (MR)

Pathophysiology—When the mitral valve leaks, blood flows from the left ventricle in two directions: into the aorta and back into the left atrium. Left ventricular end diastolic volume increases. As left ventricular function deteriorates, left ventricular end diastolic pressure increases and cardiac output eventually falls.

Symptoms and Signs—Patients present with symptoms of decreased cardiac output such as fatigue, dyspnea, weakness, and, perhaps, cachexia. Palpitations due to atrial arrhythmias may be felt. If pulmonary vascular resistance is increased, right heart failure results. If pulmonary pressures are high the patient may complain of orthopnea, DOE, and PND. The symptoms of MR are less episodic than those of MS.

Physical examination discloses a loud murmur that may radiate to the axilla. The murmur is usually pansystolic in MR caused by rheumatic heart disease, but it may be shorter if the MR is caused by mitral valve prolapse or ischemia. The EKG reveals evidence of left ventricular and/or right ventricular hypertrophy, left atrial enlargement and, in chronic cases, atrial fibrillation. Chest x-ray may show extreme left atrial enlargement and left ventricular enlargement. Echocardiogram shows left atrial enlargement, a hyperdynamic left ventricle, failure of coaption of the mitral valve leaflets, and a regurgitant jet on color Doppler examination. Calcifications of the mitral valve may be seen on chest radiograph.

AORTIC STENOSIS

Pathophysiology—Aortic stenosis causes obstruction to the flow of blood from the left ventricle. Cardiac output is maintained by the generation of increased pressures by the left ventricle. The left ventricle responds to this situation by developing concentric hypertrophy without dilatation if the obstruction develops gradually. The diameter of the normal aortic orifice is 3 to 3.5 cm^2, and a reduction to 0.5 to 1.0 cm^2 is critical.

The patient initially develops symptoms during exercise because cardiac output cannot be increased to meet the oxygen demands of exercise. Later, as the left ventricle begins to fail, cardiac output cannot be maintained at rest.

Symptoms and Signs—Aortic stenosis may exist for years before symptoms develop. The onset of symptoms for rheumatic aortic stenosis is usually in the 4th or 5th decade. The characteristic symptoms are fatigue, exertional dyspnea, angina, and syncope. The syncope is usually exertional and occurs when cardiac output cannot be increased. Reduced cerebral blood flow may cause syncope. An arrhythmia also may result in decreased cardiac output and syncope. Very late in the course of the disease, the patient has the symptoms and signs of left ventricular failure and, finally, in the preterminal phase, symptoms and signs of right heart failure.

When aortic stenosis occurs with mitral stenosis, less blood fills the left ventricle so less of a pressure gradient develops across the aortic valve. The left ventricle does not hypertrophy as much, and less angina occurs. When aortic stenosis occurs with mitral stenosis, the patient has more of the symptoms and signs of mitral stenosis.

On physical examination of the patient with aortic stenosis, an ejection click may be heard as well as closure of the aortic valve if the valve is not calcified The pulmonary valve may close before the aortic valve resulting in paradoxical splitting of the second heart sound. A systolic ejection murmur, which begins after the first heart sound, increases in intensity reaching a peak in the middle of the ejection period and decreases in intensity until closure of the aortic valve is heard. The ejection murmur thus is referred to as crescendo decrescendo.

Once the patient has become symptomatic, the prognosis is poor with 80% mortality at 4 years. Congestive heart failure accounts for mortality in up to two-thirds of the patients, and its onset suggests an average prognosis of 1 1/2 years. Ten to 20% of the patients die from an arrhythmia.

AORTIC INSUFFICIENCY (AI)

Pathophysiology—In AI a fraction of the stroke volume flows retrograde into the left ventricle so that cardiac output decreases. To compensate for the decrease in cardiac output, left ventricular end diastolic volume increases to allow for greater stroke volume. The left ventricle dilates to accommodate the increased end diastolic volume. Eventually, left ventricular function fails, and cardiac output decreases.

Symptoms and Signs—A patient who develops AI is usually asymptomatic for 10 to 20 years. The first symptom is an uncomfortable awareness of the heart beat particularly in the supine position or during exertion or emotional upset. Next, exertional dyspnea develops as a sign of decreased cardiac reserve. Later, signs of left ventricular failure appear. The patient may complain of chest pain due to pounding of the chest wall. Typical or atypical angina may develop, may be prolonged, and will not respond to nitroglycerin. Finally, symptoms and signs of systemic fluid overload and right heart failure appear. The cause of death may be pulmonary edema. Syncope is rare.

Physical examination of patients with AI reveals an increased systolic pressure and decreased diastolic pressure with a wide pulse pressure. A diastolic high-pitched blowing decrescendo murmur is heard. The murmur becomes louder and longer as the Al worsens. EKG shows left ventricular enlargement. Chest radiograph shows left ventricular enlargement and dilatation of the ascending aorta. Echocardiography shows left atrial and left ventricular enlargement and high frequency fluttering of the mitral valve.

The prognosis in decompensated AI is poor. Surgical correction is necessary before left ventricular deterioration occurs.

TRICUSPID STENOSIS

Pathophysiology—Tricuspid stenosis presents an obstruction to outflow from the right atrium and results in an increased end-diastolic pressure in the right atrium. The increased right atrial pressure causes backup of blood and congestion in the systemic circulation. Cardiac output decreases because of decreased return to the left atrium.

Symptoms and Signs—The patient presents with the symptoms and signs of right heart failure. A diastolic murmur is characteristic of tricuspid stenosis.

TRICUSPID INSUFFICIENCY

Pathophysiology—Some blood from the right ventricle flows back into the right atrium leading to enlargement of the right atrium and increased right atrial pressure. The increased right atrial pressure leads to systemic venous congestion.

Symptoms and Signs—The patient with advanced tricuspid insufficiency exhibits the signs of right heart failure and decreased cardiac output. A blowing systolic murmur is heard in tricuspid insufficiency. Atrial fibrillation may be present.

Disorders of Cardiac Rhythm (Electrophysiology)

Dysrhythmias ("arrhythmias") are irregularities in the heart rhythm that result from disturbances in the generation or conduction of the impulse. Certain dysrhythmias occur in the absence of any detectable disease of the heart. Other dysrhythmias occur characteristically in certain diseases of the heart or with toxic amounts of drugs. Predisposing factors for the development of a dysrhythmia include ischemic heart disease, congestive heart failure, hypoxemia, electrolyte imbalance, acidosis, and treatment with certain drugs such as sympathomimetics or antiarrhythmic drugs. The treatment of a dysrhythmia may be difficult unless all predisposing factors are corrected.

Normal Physiology—The conduction of an impulse through the myocardium proceeds in an orderly fashion so that both atria contract together shortly before both ventricles. The heart rate normally is controlled by the SA node, which fires at 60 to 100 beats/min. The electrophysiology of the pacemaker dictates that the faster pacemaker controls the heart rate: SA node 60 to 100/minute; AV node 40 to 60/minute; ventricular pacemaker 20 to 40/minute.

An impulse is conducted from the SA node through the atria to the AV node. Atrial depolarization causes the P wave of the EKG. The AV node slows the impulse so that the atria may contract to fill the ventricles. The impulse then proceeds down the common bundle of His, the bundle branches and into the Purkinje fibers. The QRS complex of the EKG is caused by ventricular depolarization and the T wave is caused by ventricular repolarization.

Pathophysiology—Many dysrhythmias result from a decrease or increase in automaticity of myocardial tissue. Increased automaticity may result from a more-rapid rate of depolarization, more-negative

threshold potential, less-negative resting potential, or a combination of these alterations. A decreased automaticity results from the opposite situations. Conduction disturbances, particularly slowing or failure of propagation, also may cause dysrhythmias. Conduction disturbances are caused electrophysiologically by low resting potential, a slowly rising action potential and delayed recovery from depolarization.

Many paroxysmal tachycardias are due to reentrant phenomena (ie, a circus movement) in which an impulse is propagated continually in a circuit of excitable tissue. Such circuits may exist because of structural abnormalities, such as a bypass tract, or because of functional abnormalities of diseased heart tissue. When a critically timed impulse comes to two potential pathways with different refractory periods, it may be blocked down one pathway but conducted down the second. The impulse can then be conducted up the initially refractory pathway in a retrograde direction and back down the second pathway, thus setting up the circus movement.

Dysrhythmias may have various or no effects on the individual. Significant changes in the heart rate may impair cardiac output. In pathological bradycardia, cardiac output fails to increase during conditions of increased demand such as exercise, infection, or stress. In tachycardia the synchrony of atrial-ventricular contraction may be lost or the time for ventricular filling may be decreased so that cardiac output is decreased.

Heart rate is a determinant of myocardial oxygen consumption. Coronary artery blood flow to the ventricles occurs only in diastole. Tachycardia increases cardiac oxygen demand while decreasing supply.

Pathophysiology and Symptoms and Signs of Common Arrhythmias—Sinus bradycardia is a heart rate of less than 60 beats/minute with the impulse originating in the sinus node. Sinus bradycardia occurs in individuals who are in excellent physical condition, have increased parasympathetic tone, intracerebral pressure or hypothyroidism, or in patients with SA node dysfunction due to degenerative or ischemic heart disease.

Sinus arrest refers to total cessation of sinus node activity. This may occur because of complete sinoatrial block (interference of conduction between sinus node and atrium) or loss of automaticity. There is a pause of at least 3 seconds between two P waves on the EKG. Causes of sinus arrest include excessive vagal stimulation, ischemic heart disease, and digitalis toxicity.

Sinus arrhythmia usually is not a dysrhythmia but a normal change in heart rate (less than 10% variation in length of adjacent sinus cycles) that occurs with respiration. Heart rate increases during inspiration and decreases during expiration.

Sinus tachycardia is a heart rate of greater than 100 beats/minute with the impulse originating in the sinus node. Usually sinus tachycardia is less than 140 beats/minute. The etiologies of sinus tachycardia include anxiety, fever, anemia, blood loss, thyrotoxicosis, pregnancy, pheochromocytoma, hypoxemia, various drugs, and electrolyte disturbances.

Premature atrial depolarizations (PADs) are ectopic atrial beats. Usually PADs are of little significance, although they may precede a more serious atrial arrhythmia. The rhythm with PADs is usually irregular. The P wave is abnormal in PADs or may be hidden in the T wave. A PAD may be confused with a premature ventricular contraction. The etiology of PADs is related to stimulation by nicotine, caffeine, or sympathomimetics or the deranged electrophysiology of failing atria.

Paroxysmal supraventricular tachycardia (PSVT) is a sudden attack of atrial tachycardia that is sustained by reentry. The heartbeat is regular and 140 to 250 beats/minute. This is a benign arrhythmia unless the rate is very rapid. PSVT occurs in young people with no obvious cardiac disease, and the precipitating event is usually emotional upset, trauma, fatigue, indigestion, stimulant drugs, or alcohol ingestion. The patient may become very anxious because of prominent palpitations. PSVT may end abruptly, spontaneously, or be terminated by carotid massage or medications. The prognosis for PSVT is excellent unless the rapid rate results in CHF, angina, or myocardial infarction.

Atrial flutter is a regular rhythm with an atrial rate of 250 to 350, usually 300 beats/min. The ventricular rate is 75 to 150 beats/minute reflecting AV block. The rhythm is sustained by reentry. The EKG shows sawtooth flutter waves instead of P waves. Atrial flutter occurs in patients with ischemic heart disease, mitral stenosis, thyrotoxicosis, hypertension, atrial septal defect, and hypoxemia due to chronic lung disease.

Atrial fibrillation is an arrhythmia in which the atria do not contract. The atrial rate is 400 to 600, and the ventricular rate is 80 to 180. The ventricular rate, which is slower than the atrial rate because of AV block, is usually rapid and irregularly irregular. The EKG shows fibrillating undulations instead of P waves. Because the atria do not contract, cardiac output is decreased, and the symptoms and signs of congestive heart failure may be seen. Blood stagnates in the fibrillating atria, and thrombi may form and embolize to either the lungs or the systemic circulation. The patient also may complain of palpitations due to the irregular rhythm. Paroxysmal atrial fibrillation may precede the onset of permanent atrial fibrillation in patients with mitral stenosis, constrictive pericarditis, ischemic heart disease, CHF, and thyrotoxicosis.

Premature ventricular depolarizations (PVDs) are beats that originate in an ectopic ventricular pacemaker. No P waves precede the QRS complex, which appears widened and bizarre. PVDs are a benign arrhythmia when they occur in young people without underlying heart disease. The precipitating factors in these individuals include the consumption of caffeine, nicotine or alcohol, emotional stress, and reflexes from the GI tract. PVDs may be a more serious arrhythmia when their frequency increases, they occur in pairs or runs, occur on the T wave, or originate from multiple foci. In these cases, the arrhythmia may precede ventricular tachycardia or ventricular fibrillation. PVDs are associated with ischemic heart disease, myocardial infarction, and digitalis intoxication.

Ventricular tachycardia is an arrhythmia with a rate of 150 to 250 and a regular rhythm. The rhythm originates from an ectopic focus or occurs by a reentrant mechanism. The P wave is often independent of the QRS complex (A-V dissociation). Cardiac output is decreased markedly, and the patient is usually unconscious if the arrhythmia is sustained. Ventricular fibrillation may originate from ventricular tachycardia. Common causes of ventricular tachycardia are an acute myocardial infarction, chronic ischemic heart disease, digitalis, and Type 1 antiarrhythmic drug toxicity. Ventricular tachycardia rarely occurs in a healthy individual.

Ventricular fibrillation is an irregular chaotic rhythm that is associated with no cardiac output and death if the arrhythmia is prolonged. The EKG reveals disorganized waveforms.

Abnormalities of Conduction

Normally the AV node delays the impulse from the atria. In pathologic conditions, the impulse may be delayed abnormally or blocked completely.

First-degree (1°) heart block is a dysrhythmia that usually requires no treatment. In 1° heart block, the delay of atrial impulses by the AV node is prolonged (PR interval is greater than 0.20 sec). Each atrial impulse is conducted through the AV node and results in a ventricular impulse. First-degree heart block can result from any inflammatory or degenerative disease of the heart, ischemic heart disease, and multiple medications. In healthy persons, an increase in vagal tone may result in 1° heart block.

Second-degree (2°) heart block is a dysrhythmia in which the atrial rate is greater than the ventricular rate. Mobitz Type 1 (Wenckebach) is progressive lengthening of the PR interval until an atrial impulse is not conducted to the ventricle and the corresponding QRS does not occur. The dropped ventricular beat may occur after every 2nd beat, 2:1 block or less frequently. The block may disappear during exercise or with a decrease in vagal stimulation. The atrial rate is regular while the ventricular rate is irregular. Mobitz Type 1 block is caused by ischemic heart disease, disease that involves the AV node, and by increases in vagal tone. The dysrhythmia requires no treatment unless cardiac output is impaired.

Mobitz Type 2 is a more serious block of the lower His bundle complex that may progress to complete heart block. The EKG shows a normal or increased PR interval that remains constant. QRS complexes may be dropped after the P wave on a 2:1, 3:1, 4:1, or irregular basis. Mobitz Type 2 block occurs in myocardial infarction, chronic ischemic heart disease, myocarditis, and in sclerosing diseases of the myocardium.

Complete or third degree (3°) heart block involves a normal P wave that is unrelated to QRS complexes. The atrial impulse is blocked completely from conducting into the ventricles, and the CO is maintained by the ventricles' own pacemakers. Digitalis toxicity, myocardial infarction, and degeneration of the conduction tissue cause third degree heart block. The prognosis for 3° heart block depends on whether the patient is symptomatic and the exact site of the block. Treatment is by pacemaker insertion.

Syncopal episodes due to bradydysrhythmias with resultant decreased CO are known as the Stokes-Adams-Morgagni syndrome.

Hypertension

This means abnormally elevated blood pressure. It may refer to increased pressure in any blood vessel, such as pulmonary or portal hypertension. However, it usually refers to an elevated systemic arterial blood pressure. Hypertension is not a disease but is a physical finding. Hypertension is defined as a systolic blood pressure (SBP) of greater than 140 or a diastolic blood pressure (DBP) of greater than 90 mmHg: an elevation of either SBP or DBP defines the presence of hypertension. Prehypertension, SBP 120-139 or DBP 80-89, may also merit treatment, because above 120/80 the frequency of complications due to hypertension rises significantly. Isolated systolic hypertension (SBP $>$ 140 with DBP $<$ 90) also may occur, particularly in the elderly. Above the age of 50, increased cardiovascular complications are more strongly associated with elevations of SBP than of DBP.

Normal Physiology—Blood pressure is determined by cardiac output and peripheral resistance (BP = CO $\times$ PR). Cardiac output is determined by stroke volume and heart rate (CO = SV $\times$ HR). Vascular resistance is inversely proportional to the 4th power of the internal radius of the blood vessels, according to the law of Poiseuille ($R \propto \frac{\text{length}}{\text{r4}}$). Therefore, variations in the internal lumen of blood vessels profoundly affect the blood pressure. Blood pressure varies throughout the day in any individual and is affected by physical activity, emotional upset, and other factors.

Epidemiology—Approximately 20% of the population in the US has hypertension. The incidence depends on age, race, and gender. For example, Blacks at any age have twice the incidence of hypertension as Caucasians. Hypertension is slightly more common in males than in females.

Etiology—Between 5 and 10% of the cases have an identifiable cause, and these are called secondary hypertension. Causes include renal disease, renovascular disease, endocrine disorders, and coarctation of the aorta, which will be discussed below. The remaining 90 to 95% of the cases have no known cause and are called idiopathic, primary, or essential. The etiology of essential hypertension is probably multifactorial and may involve a number of abnormalities in physiological regulatory systems. The pressure receptors in the cardiovascular system, ie, baroreceptors, may become reset at a higher pressure in response to chronic stress, overactivity of the sympathetic nervous system, or heredity. The kidneys may retain too much salt and water in response to alter reflexes that tend to maintain abnormal intravascular fluid volumes. Statistical evidence correlates the incidence of hypertension with the quantity of dietary sodium chloride. Approximately 30% of patients with hypertension are salt-sensitive; their blood pressure falls significantly with salt restriction.

Systolic hypertension most commonly occurs in elderly individuals and diabetics with stiff, noncompliant blood vessels, ie, atherosclerosis. Systolic hypertension also occurs in situations of increased cardiac output such as anemia, fever, beriberi, aortic valve insufficiency, arteriovenous fistulas, and thyrotoxicosis.

Pathophysiology—Hypertension is a major risk factor for atherosclerosis and cardiovascular complications such as CHF, AMI, and angina pectoris (see previous discussions). Sustained hypertension results in damage in the target organs: the eyes, brain, heart, and kidneys.

Damage to the eyes has been classified by Keith, Wagener, and Barker. Grades I and II retinopathy correlate well with duration of hypertension, while Grades III and IV correspond to severity.

Grade I: Arteriolar narrowing with mild depression of the venule by the crossing arteriole. Grade II: Greater arteriolar narrowing and compression of the venule by the crossing arteriole (AV nicking). Grade III: Arteriolar spasm, hemorrhages, and exudates. Grade IV: all other findings, plus papilledema. Hypertensive retinopathy leads to visual disturbances.

Damage to the brain results from cerebral edema, thrombosis, and hemorrhage (see discussion of strokes). Strokes are 12 times more common in hypertensive patients. The stroke may be small and result in focal signs or a large, fatal, cerebral hemorrhage.

The heart compensates for the increased work imposed by the increased afterload with left ventricular hypertrophy. Eventually, left ventricular function deteriorates; the chamber dilates and left ventricular failure occurs (see previous discussion of heart failure). Mortality from hypertensive congestive heart failure is 50% in 5 years. Hypertension accelerates coronary atherosclerotic heart disease and increases myocardial oxygen consumption. Angina pectoris and myocardial infarction are more common in hypertensive patients (see previous discussion of CAD).

Hypertension causes intimal and muscular hypertrophy of afferent arterioles. Malignant hypertension causes fibrinoid necrosis in the afferent arterioles and rapid deterioration if renal function. Eventually renal failure occurs (see the discussion of chronic failure below).

Symptoms and Signs—Hypertension per se causes no symptoms or signs unless the BP is very high. The symptoms and signs of essential hypertension are secondary to target organ damage. For example, retinopathy causes scotomas, blurred vision, and finally, blindness. In severe accelerated hypertension, CNS symptoms may include lethargy, confusion, increased neuromuscular irritability, convulsions, and coma. Damage to the heart results in angina pectoris or the symptoms and signs of CHF or AMI. The symptoms and signs of chronic renal failure are described later.

Elevated blood pressure may be an incidental finding during routine physical examination. The diagnosis of hypertension is based on documentation of increased blood pressure on several independent readings unless target-organ damage is already present. Adequate treatment of hypertension reduces its mortality and morbidity.

SECONDARY HYPERTENSION—This presently accounts for only 5–10% of the cases. It may be cured or ameliorated if the underlying disorder is treated successfully.

Renal vascular hypertension is mediated by the renin-angiotensin system. Renal blood flow is decreased by renal artery stenosis secondary to fibromuscular dysplasia or atherosclerosis. The renal artery lesion may be either unilateral or bilateral. The decreased renal blood flow is sensed by the juxtaglomerular apparatus, which secretes renin. Renin cleaves angiotensinogen to angiotensin I (a decapeptide). Converting enzyme in the pulmonary circulation converts angiotensin I to angiotensin II. Angiotensin II (an octapeptide) constricts blood vessels and stimulates aldosterone production. Aldosterone stimulates the retention of sodium and water and the excretion of potassium by the distal convoluted tubule. Renal parenchymal disease also is associated with hypertension; the mechanism is not understood well. The decreased excretion of sodium and water that occurs in renal failure results in volume expansion that contributes to hypertension.

Endocrine disorders cause hypertension by the production of hormone by tumors or hyperplasia of endocrine glands. Hypertension is seen in Cushing's syndrome, primary hyperaldosteronism, and hyperparathyroidism (see later discussion of these disorders). A very rare cause of hypertension is a tumor of the adrenal gland known as pheochromocytoma, which secretes excessive quantities of norepinephrine and epinephrine. Elevations of blood pressure are often episodic. Accompanying symptoms of excessive catecholamines include acute pounding headache, tachycardia, and sweating. Administration of oral contraceptives can rarely cause hypertension. Estrogens increase hepatic synthesis of renin substrate and angiotensin-I. The hypertension reverts to normal when the oral contraceptives are discontinued.

Coarctation of the aorta is a congenital malformation of the aorta exulting in a narrow area in the aorta, usually in the arch. Alterations in hemodynamics lead to a decreased renal blood flow, which activates the renin-angiotensin system.

PULMONARY DISEASES

Normal Physiology

Respiration involves all the processes in the transfer of oxygen from the air to the mitochondria of cells and of carbon dioxide from the cells back to the air. Four major steps are involved in respiration: ventilation, alveolar diffusion, transport, and tissue diffusion.

Ventilation is the functioning of the lungs to move air in and out to maintain the appropriate concentrations of oxygen and carbon dioxide in the alveoli. The process of ventilation requires proper functioning of the respiratory center in the brainstem, the peripheral nerves to the muscles, the muscles such as the diaphragm, the intercostals, the abdominals and others, and the lungs themselves. Spirometry is a technique that is used to measure the ventilatory functioning of the lungs.

For purposes of measuring lung function, the lung is divided arbitrarily into various volumes and capacities. Tidal Volume (TV) is the amount of air moved in and out of the lungs during a normal breath. The amount of air remaining in the lungs after

a maximal exhalation is called the Residual Volume (RV). The level to which the lung volumes return after a normal breath is called the Functional Residual Capacity (FRC). If one takes a maximal inspiration, filling the lungs with as much air (or gases) as possible, one then reaches the Total Lung Capacity (TLC). The Vital Capacity (VC) is the maximum amount of air that can be exhaled following a maximal inspiration. The VC represents the ability of the subject to change the size of the thoracic cavity; ie, the bellows function of the lung. Age, sex, size, and disease may affect the vital capacity. When the vital capacity is forcibly exhaled, the measurement is called the Forced Vital Capacity (FVC). The rate of exhaling the FVC is measured at time intervals, ie, Forced Expiratory Volume in one second (FEV_1), FEV_2, etc. The volume exhaled during the timed interval may be expressed as percentage of the vital capacity (FEV_1/FVC). This value is useful in assessing the severity of obstructive airway disease. The measurement of the airflow during the middle 50% of the VC is relatively independent of patient effort and is useful in determining the mechanical properties of the lung. This is called the Forced Expiratory Flow (FEF) from 25% to 75% of the VC, (FEV25–75%). This measurement is a sensitive spirometric measurement for the detection of early obstructive lung disease.

Each breath contains a portion of air that does not come in contact with a gas-exchanging membrane, such as the air in the large conducting airways. This is called dead space. The larger the dead space, the smaller the proportion of each breath which reaches a gas exchanging membrane and this affects the alveolar and arterial content of oxygen (O_2) and carbon dioxide (CO_2). In a steady state, the amount of CO_2 eliminated from the lung per minute is equal to the amount of CO_2 produced by the body. Since the partial pressure of CO_2 in the artery ($PaCO_2$) is almost equal to the partial pressure of CO_2 in the alveoli ($PACO_2$), the measurement of $PaCO_2$ assesses the adequacy of alveolar ventilation. An elevated $PaCO_2$ (>42 mmHg) means alveolar hypoventilation and a decreased $PaCO_2$ means alveolar hyperventilation.

Alveolar Diffusion—Gases are exchanged across the alveolar-pulmonary capillary membranes. The ability for this diffusion to occur depends on (1) the surface area of the alveoli, (2) the gradient between the partial pressures of gases in the alveoli and those in the blood, (3) the condition of the membranes, and (4) the amount of hemoglobin in the red blood cells. When a person breathes 100% oxygen, the gradient between the partial pressure of O_2 in the alveoli and that in the blood is so great the oxygen reaches the blood very rapidly regardless of reduction in surface area, changes in diffusion, or decreases in hemoglobin concentration. Under normal circumstances, the partial pressure of oxygen in the arteries (PaO_2) approximates the partial pressure of oxygen in the alveolus (PAO_2). The difference in these measurements, the alveolar-arterial oxygen gradient ($P(A\text{-}a)O_2$) is a measurement of the efficiency of the lungs in transferring oxygen into the blood. A normal $P(A\text{-}a)O_2$ is 10 to 15 mmHg in young people. This value increases with age.

Transport in the Blood—The maximum amount of oxygen that the blood can carry is called oxygen capacity and is determined by the amount of hemoglobin in the blood. One gram of hemoglobin can carry 1.39 ml of oxygen. The presence of hemoglobin increases the oxygen-carrying capacity of the blood by 30- to 100-fold. Normally 97% of the oxygen is carried bound to hemoglobin. The actual amount of oxygen carried, which is usually less than the oxygen capacity, is the oxygen content. The oxyhemoglobin saturation (SaO_2) is the O_2 content divided into O_2-carrying capacity × 100 and is expressed as a percentage. The oxygen content can be calculated from the oxygen saturation and the hemoglobin content. The best measurement of tissue oxygenation is O_2 delivery, which is the product of cardiac output and O_2 content. A patient with normal lungs but with extremely low hemoglobin would have a normal PaO_2 because the amount of O_2 dissolved in the plasma would be normal but the blood actually would be carrying little O_2 to the tissues because of the decreased carrying capacity. Also, the oxygen-carrying capacity of hemoglobin may be affected by physiological conditions that change the pH or temperature of the blood.

Hypoxemia

This refers to decreased amounts of oxygen in the arterial blood. There are five general mechanisms for its development.

Low Inspired-Oxygen Tension—This is not a disease but the result of the person breathing air that has less than the normal amount of oxygen. Such conditions exist at high altitudes and in some deep mines where methane may replace oxygen. As long as the lungs are normal, the $P(A\text{-}a)O_2$ will be normal. Ventilation remains normal or may be increased so the elimination of CO_2 is normal or increased.

Primary Hypoventilation—This condition occurs when the lungs no longer move air in and out to maintain appropriate concentrations of gases. The lungs themselves may or may not be normal. Primary hypoventilation may be caused by abnormalities in the respiratory centers, the peripheral nerves to the muscles, the muscles of respiration, or the chest wall. If the lungs are normal, the PaO_2 will be essentially normal but the $PaCO_2$ will be increased indicating inadequate alveolar ventilation. Drugs suppressing the ventilation centers are probably the most-common cause of primary hypoventilation.

Mismatching of Ventilation to Perfusion (V/Q Abnormalities)—If each alveolus were perfused with the appropriate amount of blood for maximum gas exchange, the ventilation-to-perfusion ratio would equal one. Normally, in the erect position, there is excess ventilation to perfusion in the apices of the lung and excess perfusion to ventilation at the lung bases. At the apex V/Q = 3 and at the bases V/Q = 0.6. In normal individuals the overall V/Q = 0.8. Airflow obstruction decreases ventilation while perfusion remains unchanged. In this situation the V/Q ratio is less than normal. If blood flow to an area is restricted while ventilation remains normal, the V/Q ratio is very high. When no ventilation is present but perfusion is normal, V/Q = 0; this is defined as a true shunt. When there is no perfusion but ventilation is normal, V/Q = ∞ _, this is defined as dead space. High V/Q ratios do not decrease the PaO_2 as ventilation is more than adequate to supply O_2 to the capillaries, which have decreased blood flow. However, low V/Q ratios do characterize hypoxemia, as ventilation is inadequate to oxygenate the relatively increased blood flow to that area. V/Q mismatching, which is the most common cause of hypoxemia, may be corrected by allowing the patient to breathe 100% oxygen for 10 to 15 minutes. This is because replacing nitrogen (which is normally 79% of the gas in the alveolus) with oxygen raises alveolar PO_2. Also V/Q mismatching results in an increased $P(A\text{-}a)O_2$. Low V/Q ratios, which occur normally at the lung bases, probably account for much of the normal $P(A\text{-}a)O_2$. Chronic bronchitis, emphysema, asthma, and many other lung diseases cause hypoxemia by affecting ventilation and lowering the V/Q ratios in many areas of the lung.

True Right-to-Left Shunting—This occurs when venous blood goes from the right heart through the pulmonary circulation without contacting a gas-exchanging surface (ventilated alveolus). Such a situation exists in pulmonary arteriovenous malformations where the pulmonary capillaries are bypassed, in the adult respiratory distress syndrome, in atelectasis where alveoli are airless, and in pneumonia and pulmonary edema where the air in the alveoli is replaced by fluid. Since the blood is not in contact with an alveolar membrane that can exchange oxygen, breathing 100% oxygen will not correct hypoxemia that results from right-to-left shunting.

Diffusion Defects—Diffusion defects are caused by thickened alveolar membranes. This is not a cause of significant hypoxemia in a resting patient but probably does play a role during exercise. Breathing 100% oxygen may increase the gradient across the alveolar membrane sufficiently to overcome a diffusion defect.

Airflow Obstructive Disease

Obstructive disorders, the most common diseases of the lungs, are characterized by an increase in airway resistance. Alterations in resistance may be acute or chronic, reversible or irreversible.

CHRONIC BRONCHITIS—Chronic bronchitis is a disease associated with excessive tracheobronchial mucus production sufficient to cause daily cough with expectoration of sputum for at least 3 months/year for 2 consecutive years. Chronic bronchitis is a clinical diagnosis that is made after other pulmonary diseases are excluded. Emphysema is defined as distention of the airspaces distal to the terminal bronchioles with destruction of the alveolar septa. The diagnosis of emphysema is based on anatomical alterations and frequently is made at autopsy. However, the entity can be considered to be present on the basis of certain physiological studies. These two diseases, although distinct processes, are often present simultaneously.

Etiology—The etiologies of these diseases have not been delineated clearly, although a variety of host and environmental factors have been implicated. Respiratory infections with viruses, *Mycoplasma*, and bacteria may play a role in the development of chronic bronchitis. Cigarette smoking correlates with the prevalence and severity of chronic bronchi-

tis and emphysema and is by far the most common cause. Currently, these diseases occur more commonly in males over 35 years, although the incidence in females is increasing, paralleling the increase in cigarette smoking by women. Air pollution has been incriminated in the etiology of both chronic bronchitis and emphysema. Also, people who work in occupations associated with dusts and noxious gases have a higher incidence of chronic bronchitis. The hereditary deficiency of the enzyme alpha-1-antitrypsin is associated with the development of severe emphysema relatively early in life in both men and women.

Pathology—Chronic bronchitis is associated with hyperplasia and hypertrophy of the mucus-producing glands in the large airways. In the small airways, there is goblet-cell hyperplasia, mucosal and submucosal inflammation and edema, peribronchial fibrosis, and intraluminal mucus plugs. Ciliated cells are lost. Emphysema is classified according to the pattern of involvement distal to the terminal bronchioles. Centrilobular or centroacinar emphysema involves the respiratory bronchioles. Panacinar emphysema involves the respiratory bronchioles, the alveolar ducts, the alveoli, and their blood supply. Both forms of emphysema often occur in a single patient, although one form may predominate.

Pathophysiology—Both chronic bronchitis and emphysema can exist without clinically significant airflow obstruction. However, using sophisticated pulmonary function testing, early disease can be detected in young smokers. Both diseases result in narrowing of the airways with increased airway resistance and decreased FEF rates. Due to the altered pressure-airflow relationships, the work of breathing is increased in chronic bronchitis and emphysema. In both diseases, the TLC and RV are increased. The hypoxemia results from ventilation to perfusion mismatching. The $PaCO_2$ may be normal, be decreased because of hyperventilation, or be elevated in severe disease or during an acute exacerbation. The chronic hypoxia leads to pulmonary vascular constriction and pulmonary artery hypertension. The chronic increased afterload on the right heart ultimately leads to right heart failure (cor pulmonale). Other sequelae of severe hypoxemia include polycythemia and alteration of the patient's mental status

Symptoms and Signs—Dyspnea on exertion and functional disability result from severe airway obstruction with its increased work of breathing.

PREDOMINANT EMPHYSEMA—These patients have a long history of exertional dyspnea with little cough or sputum production. The typical patient is thin, uses accessory muscles to breathe, is tachypneic, with prolonged expiration through pursed lips, frequently leans forward when sitting, has a hyperresonant percussion note, and has diminished breath sounds by auscultation. The chest radiograph reveals low and flattened diaphragms and signs of hyperinflation. The clinical course is progressive, severe dyspnea for which little can be done. Resting blood gases become abnormal late in the course of the disease.

PREDOMINANT BRONCHITIS ALONG WITH EMPHYSEMA—The typical patient has an impressive history of cough and sputum production for many years. Acute exacerbations increase in frequency, duration, and severity over the years. After each episode, the patient's baseline status may have deteriorated slightly. The presenting complaints may include cough, sputum production, exertional dyspnea, or peripheral edema secondary to right heart failure. This patient is usually overweight, cyanotic, and only slightly tachypneic. On auscultation coarse rhonchi and wheezes may be heard throughout the lung fields. Arterial blood gas analysis reveals hypoxia and hypercapnia. The VC is normal or only slightly decreased while the FEF rates are low. Some of these patients develop emphysema with the resultant symptoms. A patient with chronic bronchitis may experience many episodes of acute respiratory failure usually precipitated by a respiratory tract infection.

Reversible Airway Obstruction

BRONCHIAL ASTHMA OR REACTIVE AIRWAYS DISEASE—This is defined as a disease characterized by increased responsiveness of the trachea, bronchi, and bronchioles to various stimuli and is manifested by widespread narrowing of the airways that changes in severity either spontaneously or as a result of therapy.

Etiology and Epidemiology—Asthma affects at least 2% of the population. About one-half of the cases develop before age 10 and another third develop before age 40. Childhood asthma occurs in males predominantly (2:1), but after age 30 there is no sex difference.

Because of the diversity of the disease, the classification of asthma is difficult. Allergic or extrinsic asthma usually is found in individuals with a history or a family history of atopy or allergic diseases such as rhinitis, urticaria, and eczema. Allergic asthma, which accounts for 25% of the cases, tends to be seasonal and occurs more commonly in children and young adults. Nonseasonal allergic asthma may be due to antigens such as animal dander, molds, and dust. In another group of patients, ingestion of aspirin or nonsteroidal anti-inflammatory agents may aggravate the asthma. Asthma also may occur during times of heavy industrial air pollution, physical exercise, or emotional upset. Pulmonary infections, congestive heart failure, pulmonary embolism, and treatment with cholinergic agents or beta-adrenergic blockers may also provoke asthma. Asthma that occurs without an identifiable cause is labeled intrinsic or idiosyncratic.

Pathology—The hallmarks of acute asthma are over distention of the lungs, gelatinous plugs in the bronchioles, hypertrophy of the bronchial smooth muscle, mucosal edema, denudation of the surface epithelium, pronounced thickening of the basement membranes, and infiltration of the bronchial wall with inflammatory cells, particularly eosinophils and mast cells. Emphysematous changes are usually absent.

Pathophysiology—In those with allergic asthma, bronchoconstriction and alterations in bronchial secretions are the result of an immediate hypersensitivity reaction. In this response the interaction of antigen and antibody, particularly IgE, causes the release of chemical mediators from sensitized mast cells in the lungs. The mediators include histamine, leukotrienes, platelet-activating factor, eosinophil chemotactic factor of anaphylaxis (ECF-A), and neutrophil chemotactic factor of anaphylaxis (NCF-A). Secondary mediators include prostaglandins and bradykinin. These mediators constrict bronchial smooth muscle and increase vascular permeability.

Adenylate cyclase catalyzes the formation of the cyclic nucleotide, cyclic 3′,5′-adenosine monophosphate (cyclic AMP; cAMP), from adenosine triphosphate (ATP). cAMP is an intracellular mediator that inhibits the release of the chemical mediators. An increase in the concentration of cAMP causes relaxation of bronchial smooth muscle. It is thought that bronchoconstriction in asthmatics might result from a defect in cAMP as a result of nonresponsiveness to endogenous catecholamines due to downregulation of receptors. Catecholamines stimulate adenyl cyclase to increase the intracellular concentration of cAMP.

A second cyclic nucleotide, cyclic 3′,5′-guanosine monophosphate (cyclic GMP; cGMP), opposes the action of cAMP. Actions that are facilitated by cAMP are suppressed by cGMP and vice versa. cGMP promotes the release of bronchoconstrictor substances from mast cells. Guanylate cyclase catalyzes the synthesis of cGMP in response to stimulation by acetylcholine.

Symptoms and Signs—Symptoms include dyspnea, chest tightness, cough, and wheezing. Some patients with asthma do not wheeze and may have only dyspnea and/or cough. The symptoms are episodic and frequently occur at night. In asthma, the contraction of bronchial smooth muscle and the presence of mucosal edema and thick, tenacious mucus result in airflow obstruction. Hypoxemia is present during an acute severe attack. Blood-gas analysis usually shows decreased $PaCO_2$ and respiratory alkalosis. Normal or elevated levels of carbon dioxide during an acute episode should be viewed as impending respiratory failure. Clinical symptoms and signs are unreliable for judging tissue oxygenation. When severe symptoms persist for days or weeks, or fail to respond to basic therapy, the condition is known as status asthmaticus. Eosinophilia in sputum and blood suggests but is not specific for asthma. The chest radiograph shows hyperinflation and is not diagnostic.

Restrictive Lung Disease

This is a general term applied to a wide spectrum of diseases with a decrease in total lung capacity. In advanced cases, other lung volume components also are reduced. Most patients with restrictive lung diseases have intrinsic structural and functional abnormalities of the lung, which cause a stiff lung. Stiffness of the lungs is defined by a decrease in lung compliance or change in lung volume per unit change in pressure. A few of patients have normal lungs but have reduced lung volumes because of abnormalities of the chest wall, pleura, or abdomen.

Pathology—Although some restrictive lung diseases have unique pathology, many have similar nonspecific end-stage changes. Such changes may include pulmonary fibrosis of the alveolar septa, peribronchiolar fibrosis, mononuclear inflammatory cell infiltrate, smooth muscle proliferation within the interstitium, metaplasia of the alveolar lining cells, and vascular obliteration with pulmonary hypertension.

Etiology—Restrictive lung disease may be acute or chronic. An example of an acute, reversible restrictive lung disease is pulmonary edema. Chronic restrictive lung diseases are diverse. In asbestosis, hypersensitivity pneumonitis, drug- or toxin-induced lung injury, the etiology is known, as it is for lung disease associated with sarcoid, collagen vascular disease, or other well-defined systemic illnesses. In pulmonary alveolar proteinosis, desquamative interstitial pneumonitis (DIP), and idiopathic pulmonary fibrosis, the cause is not known.

Symptoms and Signs—The hallmark of all restrictive lung diseases is dyspnea, a sensation of shortness of breath. This results from the increased work of breathing caused by stiff lungs. In addition, airflow resistance is increased because patients breathe at low lung volumes, which allows small airways to close. Tachypnea and a nonproductive cough are common findings. Although fine crackles may be heard, auscultatory findings are usually minimal compared to the degree of pathological changes. Patients with extensive fibrosis may experience recurrent pneumothorax. Pulmonary hypertension advancing to cor pulmonale may be seen as a late sequel. This complication is caused either by obliteration of the pulmonary vascular bed or by increased pulmonary resistance due to hypoxemia.

The chest x-ray in restrictive lung diseases may show decreased lung volumes and increased interstitial markings. Arterial blood gases often reveal hypoxemia and hypocapnia.

Abnormalities on physiological testing include an increased alveolar-arterial oxygen gradient and a decreased diffusion capacity.

Adult Respiratory Distress Syndrome

This syndrome (ARDS) is a common cause of acute respiratory failure in a hospitalized patient. Its hallmark is damage to the pulmonary capillaries and alveolar epithelium leading to increased permeability and acute pulmonary edema. The etiology of this syndrome is multiple and includes shock, infection, near drowning, drug and toxin exposure, acute pancreatitis, and aspiration pneumonia. Despite the wide spectrum of diseases that may lead to ARDS, there is a similar clinical picture in all cases. Acute respiratory failure is accompanied by a diffuse infiltrate on chest x-ray and physiological disturbances of restrictive lung disease. On pathology there is edema, hemorrhage, hyaline membranes, inflammatory cells, and fibrosis.

Deep Venous Thrombosis and Pulmonary Embolism

Both deep venous thrombosis (DVT) and pulmonary embolism (PE) are significant causes of mortality and morbidity. Together they form a spectrum referred to as venous thromboembolism (VTE). The most important factor for decreasing morbidity and mortality is the prevention of DVT.

Normal Anatomy and Physiology—Veins are thin-walled vessels composed mainly of collagen with some smooth muscle and little elastic tissue. They normally contain a large proportion of the circulating blood but at significantly lower pressures than arteries. The venous system of the lower extremities is composed of the deep, superficial and communicating veins.

Blood return from lower extremities depends on the contraction of skeletal muscles, especially in the calves. Valves prevent retrograde flow of blood in the veins. These valves are present even in very small and their number decreases in the proximal veins. Valves are composed of elastic and collagen tissue and operate passively in response to pressure changes.

The lung has two arterial blood supplies. The pulmonary artery exits from the right ventricle, immediately divides into the right and left branches, and carries deoxygenated blood from the systemic venous system to the lungs for gas exchange. The bronchial arteries branch off the aorta and carry oxygenated blood to the supporting tissue of the lung.

Normally, clots do not form within the vascular system. The smooth endothelial surface of the blood vessels and a negatively charged protein layer on the endothelial surface that repels platelets are probably the most important factors in preventing clot formation. Two factors prevent excessive clotting. Approximately 85% of the thrombin formed is adsorbed to the fibrin threads, which prevents the spread of the thrombin. The remaining thrombin is inactivated in 20 min by combining with antithrombin III.

Plasma normally contains a protein called plasminogen which, when activated, forms plasmin. Plasmin is a proteolytic enzyme that digests fibrin, fibrinogen, prothrombin, and Factors V, VIII, and XII. The process that activates plasminogen is understood poorly. Plasminogen is incorporated in all blood clots and is involved in dissolution of intravascular clots.

Epidemiology—It is difficult to estimate the incidence of DVT. The incidence of PE has been estimated as high as 500,000 cases/year and is the cause of at least 50,000 deaths/year in the US. On autopsy, PE is found in 20 to 25% of deaths in general hospitals, 25% of deaths in nursing homes and as many as 50% of deaths due to congestive heart failure. The risk of VTE is increased markedly in individuals over 40 years. It is postulated that the diagnosis of PE is missed frequently in elderly chronically ill patients.

Etiology—A number of conditions and situations have been associated with increased risk of VTE. These include prolonged bed rest, immobilization, cancers (particularly adenocarcinomas of the pancreas, lungs, or prostate), polycythemia vera, congestive heart failure, administration of estrogens, the postpartum state, orthopedic injuries, major surgery, trauma, chemical irritations, and infections. Approximately 85% of pulmonary embolic episodes are caused by DVT.

Pathophysiology—Over 100 years ago, Virchow described three factors that promote venous thrombosis: stasis, hypercoagulability, and vessel wall factors. Increased platelet adhesiveness and aggregation also may be involved.

Stasis occurs at various sites in veins. The edges of the valves cause turbulent blood flow with eddy formation and stasis. Areas adjacent to the valves and the junctions of tributaries also are areas of stasis. Dilated veins (varicose veins) or previously damaged veins may have sluggish flow and incompetent valves. Lack of pumping of the blood in the veins by skeletal muscle contraction or compression of the veins by the muscle mass may explain the increased risk of DVT during bed rest or immobilization, and part of the increased risk during surgery. In polycythemia vera the blood is viscous and prone to stasis. Congestive heart failure also may increase the stasis of blood in the lower limbs. The stasis may allow the activation of factors as well as inhibit the dilution or removal of activated factors.

Various risk factors for VTE are associated with hypercoagulable states. Cancers are thought to increase production of Factors V, VIII, IX, and XI, release tissue thromboplastin from necrotic tumor, and decrease the efficiency of the fibrinolytic system. Trauma and surgery may increase plasma concentration of fibrinogen and procoagulants, increase platelet adhesiveness, and decrease fibrinolysis. Estrogens increase the production of Factors I, II, VIII, IX, and X, increase platelet adhesiveness, and decrease the activity of antithrombin III. Estrogens also dilate veins and promote stasis. Congenital abnormalities of the coagulation cascade predispose strongly to VTE, often recurrent. These include the Factor V Leiden mutation, the prothrombin 20201A mutation, and congenital deficiencies of protein C and protein S.

Increasing age predisposes individuals to thrombosis because of increased stasis caused by venous dilatation, malfunction of venous valves, decreased skeletal muscle mass, decreased physical activity, and decreased cardiac output. Increased Factor VIII and decreased antithrombin III activity enhance coagulation.

If the vessel wall is disrupted, collagen is exposed and/or tissue thromboplastin is released. Exposed collagen and the extrinsic system activate the intrinsic coagulation system via tissue thromboplastin. Platelets adhere to the exposed collagen, aggregate to form a platelet plug and release platelet Factor III. Platelet Factor III is similar to tissue thromboplastin in that it initiates the extrinsic coagulation system. Platelet Factor III also can activate Factors VIII, IX, Xl, and XII. The end product of coagulation is the thrombus, which is composed of fibrin, trapped serum, and blood cells. The clot itself initiates a vicious cycle that promotes more clotting. The clot extends until it reaches an area of faster-flowing blood.

The most-feared form of DVT involves the iliofemoral veins, since thrombi here are most likely to result in large emboli to the lungs.

When an embolus lodges in a pulmonary artery, the area is being ventilated but not perfused. The area is now dead space. The alveoli transiently constrict due to hypocapnia. Surfactant is lost and atelectasis develops in 24 to 48 hours. Hypoxemia usually develops. In massive PE, pulmonary hypertension may result and lead to acute right heart failure. Whether lung infarction occurs depends on the size of the embolus and the dual pulmonary blood flow. Many emboli are dissolved quickly by the fibrinolytic system. Recanalization may occur in 1 week. Some vessels, however, remain totally occluded with resultant loss of lung function.

Symptoms and Signs—DVT may present as swelling of the calf or thigh with edema of the lower extremity. The area over the thrombosis may be tender, warm, and erythematous. The thrombosed vein may be felt as a hard cord. Physical maneuvers of the limb or walking may worsen the pain. However, in many cases of DVT no symptoms or signs are present. More than 50% of patients with symptoms and signs normally attributed to DVT do not have VTE. The diagnosis of DVT is made conclusively by phlebography, but this invasive test is difficult to obtain

quickly, is painful, and may cause phlebitis. Noninvasive evaluation is most practicably carried out with Doppler flow studies. Elevated plasma levels of D-dimer are suggestive of VTE, and this test is becoming increasingly used. DVT can lead to PE, the postphlebitic syndrome (edema, pain, increased pigmentation, eczema, induration, and ulceration) or recurrence of DVT.

The symptoms and signs of PE depend on the size of the embolus and the presence of infarction. The classic presentation of PE is the sudden onset of dyspnea. If infarction occurs, pleuritic chest pain and hemoptysis also may be present. Hypoxemia and an increased alveolar-arterial oxygen gradient may be seen. Physical examination may or may not demonstrate the signs of DVT. Crackles, local wheezes, and a pleural friction rub may be heard on auscultation. Tachycardia and tachypnea are seen. Signs of acute right heart failure can be seen in massive PE. However, the physical examination may be completely normal. Laboratory examination is not diagnostic. The chest radiograph is often normal but may show a pleural effusion and/or infiltrate and/or changes in size or disappearance of blood vessels. A ventilation perfusion scan may give presumptive evidence for the diagnosis of PE. The test is associated with false negatives if the area involved is small or false positives if other lung diseases are present. Pulmonary arteriography is the most accurate method used to diagnose PE. CT angiography is somewhat more specific than ventilation-perfusion scanning and may even replace pulmonary angiography when the findings are definite.

Cystic Fibrosis

This is a disease with diverse clinical manifestations characterized by abnormal exocrine gland secretions. It presents in childhood; with improved methods of detection in mild cases and better treatment, more adults are followed now for this disease.

Etiology and Epidemiology—Cystic fibrosis is an autosomal recessive disease carried by a gene on the long arm of chromosome 7 that codes for cystic fibrosis transmembrane regulatory protein (CFTR), a protein with a predicted molecular weight of 170 kD. There is one common mutation, the Δ508 mutation, accounting for 70% of mutations, and over 1000 less common ones. Cystic fibrosis affects both sexes equally and occurs predominantly in Caucasians. In the past, cystic fibrosis was considered a fatal disease of childhood. With better techniques for earlier detection and improved methods of treatment, the median life expectancy has risen to 32 years.

Pathophysiology—Defects in CFTR protein, the chloride channel in the membrane of epithelial cells, impair cAMP-dependent chloride secretion by respiratory epithelium. Epithelial secretions become thick and difficult to clear. The high chloride concentration in secretions impairs bactericidal activity and predisposes to infection, particularly with Pseudomonas.

Symptoms and Signs—The initial manifestation may be intestinal obstruction in the newborn secondary to abnormally thick meconium. Early in life pulmonary complications develop. Thick, tenacious mucus results in bronchial obstruction with subsequent atelectasis and infection. The initial bacterial pathogens, including *S. aureus*, are replaced later by *P. aeruginosa* and other gram-negative organisms. Death in cystic fibrosis is usually due to overwhelming pulmonary infection and respiratory failure. With longer survivals, cor pulmonale and recurrent hemoptysis are seen more frequently.

Pancreatic insufficiency develops in approximately 80% of patients and causes malabsorption characterized by steatorrhea and deficiencies of vitamin B12, and the fat-soluble vitamins. Some patients experience recurrent bouts of pancreatitis. Biliary cirrhosis develops in approximately 10% of patients. The incidence of gallstones is increased. Most male patients are sterile because of a malformation or blockage of the vasa deferentia. Secondary sex characteristics are normal. The fertility rate among females is approximately one-fifth that of a control population. The reason for this is probably the increased viscosity of the cervical mucus.

The best initial diagnostic test is the sweat test. There is usually a 3- to 5-fold increase in the concentration of chloride in the sweat of patients with cystic fibrosis. The level of sweat electrolytes does not correlate with severity of disease. The sweat test is a difficult test to perform correctly and must be obtained in a reliable, experienced laboratory.

GASTROENTEROLOGY

Esophagus

The esophagus is a muscular, hollow tube, which extends from the pharynx to the stomach. Its major function is to transport food from the oropharynx to the stomach. It has a sphincter at both the top and the bottom end. The upper esophageal sphincter maintains a zone of high pressure between the oropharynx and the body of the esophagus. The sphincter pressure increases with respiration and prevents inspired air from entering the gastrointestinal tract. It also acts as a barrier against the regurgitation of esophageal contents into the pharynx. The lower esophageal sphincter (LES) consists of highly specialized muscles, which is tonic in the resting state. It thus maintains a zone of high pressure between the esophagus and stomach. Its major function is to prevent reflux of gastric contents into the esophagus.

The two most specific symptoms of esophageal disease are dysphagia and heartburn. Dysphagia is the sensation of food sticking in the esophagus. It always indicates esophageal disease. Dysphagia may be of two types—to solids only, indicating a mechanical disorder such as stricture or tumor, or to both solids and liquids, indicating a motility such as diffuse spasm of achalasia. Heartburn refers to a burning discomfort that typically migrates from the abdomen up the retrosternal area of the chest. Less common symptoms are chest pain and odynophagia (painful swallowing and regurgitation).

Normal Physiology—The esophagus is a muscular organ that actively transports food by means of peristaltic waves. Swallowing involves the propulsion of a bolus of food from the oropharynx through the relaxed upper esophageal sphincter. Primary peristaltic waves then transport the bolus through the esophagus and past the LES, which relaxes in response to peristalsis. Secondary peristalsis is the same as primary peristalsis, but is initiated by a bolus of material in the body of the esophagus, such as occurs with the reflux of gastric contents. Tertiary contractions are nonpropulsive, nonperistaltic waves that, for the most part, are pathologic and interfere with normal transport of food through the esophagus. Tertiary contractions are associated with dysphagia to solids and liquids and, in some patients, pain.

The regulation of esophageal function is complex and modulated by the swallowing center in the brain. Afferent impulses from the pharynx and the esophagus are mediated by the vagus nerve. The efferent impulses also are mediated vagally through cholinergic fibers splayed around the esophagus in a myenteric network known as the plexus of Auerbach. The resting tone of the esophageal body is maintained largely by cholinergic stimulation, although sympathetic innervation probably plays some regulatory role. The resting pressure of the LES is maintained by specialized circular, smooth muscle. Relaxation of the LES is mediated by a balanced cholinergic-adrenergic stimulated release of noncholinergic, nonadrenergic neurotransmitters. The resting pressure of the LES is modified by a number of factors. Factors known to increase the LES pressure are certain G-l hormones such as gastrin; foods such as a protein meal; drugs such as bethanechol, metoclopramide, erythromycin, cisapride, and domperidone; and increased gastric pH that occurs with eating. Factors known to decrease LES pressure are the GI hormones secretin and cholecystokinin; foods such as fat; certain drugs such as caffeine, alcohol, anticholinergics, calcium channel blockers, and theophylline; and a decreased gastric pH that occurs with fasting.

DISEASES OF THE ESOPHAGUS

GASTROESOPHAGEAL REFLUX DISEASE—

Pathophysiology—This is the most common disorder of the esophagus and refers to the reflux of gastric content into the esophagus with subsequent injury to the esophageal mucosa. Gastroesophageal reflux disease (GERD) is caused, in most people, by an incompetent LES such that either the resting pressure (normally 12 to 20 mmHg) is decreased or, more commonly, the LES relaxes inappropriately allowing gastric contents to reflux into the esophagus. The gastric contents (primarily acid and to some extent bile) then damage the squamous epithelium of the esophagus. In some patients, a defect in secondary peristalsis caused by smoking, or a defect in gastric emptying caused by diabetes or a gastric stapling operation, may contribute. Inflammation of the mucosa and thickening of the basal layer of epithelial cells characterize esophagitis. In some patients, erosions and ulcerations may occur. In most patients, a hiatus hernia, a bulging of the stomach into the chest cavity, occurs, but its role in the pathophysiology of GERD is thought to be relatively minor. Nevertheless, it is unusual to see severe GERD in the absence of a hiatus hernia.

Symptoms and Signs—The major symptom of GERD is heartburn, a retrosternal burning pain that migrates up the chest from the epigastrium. It is accompanied sometimes by an acid or bile taste in the back of the throat or a profusion of watery saliva (water brash). Typically' the

heartburn is aggravated by overeating, bending, straining, or lying down after eating. Dysphagia may occur with GERD, either secondary to esophageal spasm (causing liquid and solid dysphagia) or due to stricture (causing dysphagia to solids only).

Diagnosis—The diagnosis of GERD depends on the demonstration of esophagitis by endoscopy with biopsy and the demonstration of reflux of acid into the esophagus by direct measurement of pH in the distal esophagus with an esophageal pH probe. The treatment of GERD has two phases: (1) healing the esophageal mucosa, and (2) preventing recurrence. Since the injury is mediated by acid, the hallmark of therapy is acid reduction. This can be achieved with H2 receptor antagonists (cimetidine, famotidine, nizatidine, ranitidine) or proton pump inhibitors (lansoprazole and omeprazole). In general, the proton pump inhibitors (PPIs) are approximately 50% more effective for treating all grades of esophagitis to healing and are thus the treatment of choice. After 8 weeks of therapy, PPIs will heal 90–95% of patients with mild disease and 80–90% of patients with severe disease, while H2RAs heal 50% of patients with mild disease and 20% of patients with severe disease. The prokinetic drugs, metoclopramide and cisapride, despite addressing the underlying problem of lower esophageal sphincter dysfunction, do not have sufficient healing rates, have a narrow therapeutic index and are not FDA approved as primary therapy for GERD.

For most patients (ie, 90%), GERD is a lifetime disease requiring a lifetime of therapy. This can only be achieved with the proton pump inhibitors. Prokinetic drugs, because of their narrow therapeutic index and propensity for tachyphylaxis, have not been shown to maintain healing adequately. The H2RAs, because of tachyphylaxis, also do not maintain healing. To date, the only drugs shown to maintain healing above 80% are full dose proton pump inhibitors—either lansoprazole or omeprazole.

ESOPHAGEAL STRICTURE—Strictures of the esophagus may be benign or malignant. Chronic GERD or the ingestion of toxic materials such as lye usually causes benign strictures. They are manifested anatomically by a symmetric narrowing of the esophagus that can be seen either by barium swallow or esophagoscopy. They are manifested clinically by dysphagia to solids only. Malignant strictures are caused either by esophageal squamous cell carcinoma or adenocarcinoma arising from the stomach or metaplastic columnar epithelium in the esophagus (so-called Barrett's esophagus). Malignant strictures are usually irregular and asymmetric on barium swallow or endoscopy and can be diagnosed by esophageal biopsy. They usually are associated with rapidly worsening dysphagia to solids along with weight loss.

DIFFUSE ESOPHAGEAL SPASM—This is a motility disorder of the esophagus characterized by frequent and severe tertiary contractions. It occurs predominantly in elderly patients, but may be seen secondary to other disorders of the esophagus such as GERD. It is manifested clinically by intermittent dysphagia to solids and liquids and/or chest pain. Swallowing hot or cold drinks frequently precipitates the symptoms. Barium swallow or manometry makes the diagnosis.

ACHALASIA—This is a motility disorder of the esophagus characterized by an increase in lower esophageal sphincter pressure and an absence of primary peristalsis. Pathophysiologically, achalasia is caused by a loss of the myenteric plexus. This may occur as a primary defect of unknown etiology or as a secondary defect due to invasive carcinoma of the lower esophagus or infestation from Trypanosoma cruzi, the cause of Chaga's disease. The diagnosis is made by manometry, which demonstrates an increase in the LES pressure, incomplete relaxation of the LES, and a total absence of primary and secondary peristaltic waves. Tertiary contractions may be seen. There is also a characteristic x-ray appearance with the body of the esophagus dilated and tapering down to a closed esophageal sphincter (so-called "bird beak" appearance). The disorder is seen more commonly in young people in their teens and twenties, but may be seen at any age. The patients typically are afflicted with intermittent dysphagia to solids and liquids. They may have regurgitation in the supine position with choking and coughing from aspiration. Weight loss occurs as the symptoms become more severe and more continuous.

Stomach and Duodenum

The main function of the stomach is to receive ingested food and then present it to the small bowel in tiny particles suitable for digestion and subsequent absorption. The first step in this process is expansion of the stomach (so-called receptive relaxation) to accommodate the ingested liquid and solid food (chyme) without an increase in gastric pressure. The stomach then mixes, emulsifies, acidifies, and meters the chyme into the small bowel. This is achieved through gastric motility. The proximal and distal portions of the stomach have separate and distinct roles in motility. The proximal stomach receives and stores solids and is primarily responsible for the transfer of emulsified foodstuffs from the body of the stomach to the duodenum. The properties that allow this to occur are receptive relaxation (the ability to relax and receive food stuffs without increasing intragastric pressure), accommodation (the ability to distend to a large size without an increase in intragastric pressure), and contraction. The contraction waves of the proximal stomach are slow and sustained. They act to force solid meal components from the proximal to the distal portion of the stomach.

The main function of the distal stomach is to retain and grind foodstuffs and to prevent reflux of duodenal content back into the stomach. The motor activity of the distal stomach is characterized by peristaltic waves sweeping downward toward the pylorus. These contractions are lumen-obliterating such that solid particles are retropelled for further emulsification. Only when the particles are smaller than 1 mm in diameter will they pass into the duodenum. The motor function of the stomach is regulated largely by the vagus nerve.

The stomach also has a major secretory function. It secretes acid, pepsin, and intrinsic factor. The function of gastric acid is not entirely clear, but it does not play a particularly important role in digestion; rather, it seems to function more as a barrier to toxins and bacteria in the environment. It also plays a minor role in pH homeostasis. There are two types of acid secretion-basal and stimulated. Basal acid secretion occurs continuously and independently of external stimuli. It is characterized by a circadian rhythm in which acid secretion is highest from about 10 pm until midnight and lowest from about 4 am until 8 am. This pattern of acid secretion is responsible for one of the characteristic features of peptic ulcer disease, which is nighttime waking with pain when acid secretion is high and unneutralized by food.

Stimulated acid secretion, on the other hand, occurs in response to the sight, smell, and ingestion of food. This acid secretion is stimulated by acetylcholine, the neurotransmitter of the vagus nerve; the hormone gastrin, secreted by G cells in the gastric antrum and histamine, secreted by enterochromafin cells in the wall of the stomach. Acid secretion is turned off by prostaglandin E, somatostatin, and some yet to be identified enterokinase. During most of the day, the food that stimulates acid secretion also neutralizes it, keeping the pH between 4 and 5. However, when the stomach is empty, approximately 2 to 3 hours after eating, the pH again drops and ulcer patients tend to get pain that is relieved by eating or antacids.

The epithelium of the stomach, duodenum, and esophagus is protected from autodigestion by hydrochloric acid by means of a mucosal defense system. The most characteristic feature of this system is the secretion of mucous and bicarbonate. Bicarbonate is secreted by epithelial cells in the stomach and duodenum and is separated from luminal acid by a layer of mucous, which also is secreted by epithelial cells. These cells are largely under the influence of prostaglandin E_1. Thus, the net effect of prostaglandin E_1 is to decrease acid secretion and increase mucosal defense. This is another example of the adaptive or protective effects of prostaglandins in the body.

The function of the duodenum is to receive gastric contents and to mix them with secretions from the pancreas and gallbladder, which serve to digest (pancreatic enzymes) and solubilize (bile) the nutrients received from the stomach.

DISEASES OF THE STOMACH AND DUODENUM—

The major diseases of the stomach and duodenum are gastritis, gastric ulcer, duodenitis, and duodenal ulcer, all of which are in some way related to gastric acid.

PEPTIC ULCER DISEASE—Peptic ulcer disease is a spectrum of diseases consisting of gastritis, gastric ulcer, and duodenal ulcer. They

are among the most frequently encountered disorders of the gastrointestinal tract. Common to these disorders as well as gastric cancer is gastritis, an inflammation of the epithelial surface and gastric glands of the stomach. The most common cause of gastritis and thus of ulcer disease and gastric cancer, is *Helicobacter pylori* infection.

Epidemiology—Peptic ulcer disease is on the decline in the developed world, having peaked early in the century, and probably reflecting the improved sanitary conditions that reduce the spread of enteric infections such as H pylori. Nevertheless, the point prevalence is still 1%, and the lifetime incidence 10%.

Symptoms and Signs—The clinical presentation of ulcer disease is characteristic and is a reflection of the pH in the stomach. Thus, the typical burning epigastric pain occurs on an empty stomach, ie, 2–3 hours after eating and is relieved by eating. It also occurs in the late evening and early morning hours when acid secretion is high and the acid is unneutralized by eating. Ulcer disease may also present with its complications of bleeding (manifested by hematemesis, melena, or anemia), obstruction (manifested by early satiety and weight loss), and penetration/perforation (manifested by persistent epigastric pain, back pain, and fever). These are the so-called alarm manifestations of ulcer disease and should always preclude empiric treatment and dictate further work-up.

Pathophysiology—Ulcer disease occurs whenever there is an increase in acid secretion (eg, Zollinger-Ellison syndrome) or a decrease in mucosal defense (eg, non-steroidal anti-inflammatory therapy) or a combination of both (eg, *H pylori* infection). *H pylori* causes approximately 70% of ulcer disease in the developed world and more than 90% of ulcer disease in the undeveloped world. NSAIDs cause 5–10% of duodenal ulcers and 20–25% of gastric ulcers. Twenty to 30% of ulcer disease is idiopathic.

H pylori is a unique organism that is exquisitely well adapted to the gastric environment and, in fact, cannot exist outside an acidified environment. It is a gram-negative, flagellated spirochete. The flagella allow it to burrow through the mucous layer of the stomach and attach to the epithelial surface. It is a facilitative acidophile, meaning it can adjust its cytoplasmic pH to its surrounding environment. It is also microaerophilic making it highly adaptive to the interface of the oxygen-reduced environment of the gastric lumen and the oxygen-enriched environment of the gastric mucosa. Finally, it has the unique enzyme, urease, which splits urea into bicarbonate and ammonia, thus creating an alkalinized ammonia shell to interface with the acidified gastric lumen.

The pathogenesis of *H pylori* ulcer disease is only partially understood. It appears that 70–80% of ulcers are associated with *H pylori*, but that only 10% of *H pylori* infected individuals develop ulcers. Thus, host factors and co-factors are important in the pathogenesis. In general, there appear to be two patterns of infection. The first is characterized by diffuse antral gastritis that leads to increased acid secretion, secondary gastric metaplasia of the duodenum, duodenal ulcer in the gastric metaplasia, and in some patients, formation of lymphomas in the antrum. The second type of infection is a patchy atrophic gastritis involving the antrum and fundus of the stomach. It leads to gastric atrophy, decreased acid secretion, intestinal metaplasia of the stomach followed by gastric ulcer, and in some patients, gastric adenocarcinoma.

In summary, *H pylori* accounts for 70–80% of ulcers, almost 100% of gastric mucosal lymphomas and 90% of gastric cancer. The World Health Organization has classified *H pylori* as a class I (ie, definite) carcinogen and estimates that eradication of *H pylori* would lead to a 90% reduction in gastric cancer worldwide.

Diagnosis—The diagnosis of peptic ulcer disease is best made by endoscopy. Helicobacter infection can be diagnosed by gastric biopsy, a pH color indicator test based on the production of ammonium by urease in Helicobacter-infected patients or by a serum antibody test. A radioisotope test based on the urease reaction has recently been developed. In this test, the patient ingests 14C urea. If urease (ie, H pylori) is present, the urea is converted to ammonium and carbon dioxide with the 14CO2 blown off in the breath. The specificity and sensitivity of this test are both greater than 95%.

The treatment of acid peptic disease is (1) acid reduction to heal the ulcer and relieve symptoms and (2) prevention of recurrence by treating the underlying cause, either NSAIDs or *H pylori*.

Acid reduction may be achieved with H2 receptor antagonists (cimetidine, famotidine, nizatidine, or ranitidine) or proton pump inhibitors (lansoprazole, omeprazole). The proton pump inhibitors are far superior and in the case of *H pylori* disease raise the pH to sufficient levels to improve antibiotic efficacy. Either discontinuing the NSAID or increasing mucosal resistance with the prostaglandin E1 analog, misoprostil can prevent NSAID-induced ulcers. *H pylori* ulcers can be prevented by antibiotic therapy. It should be noted that *H pylori* is an organism of great genetic diversity with a high mutation rate. It is therefore important to use multiple antibiotics. It is also important to keep the gastric pH above 5 in order to create an optimum environmental pH for the antibiotics. This can only be achieved with proton pump inhibitors given twice daily. The most widely used antibiotics are metronidazole, amoxicillin, and clarithromycin. It should be noted, however, that metronidazole has a 40% drug resistance rate. The best eradication rates at the time of publication have been achieved with lansoprazole (30 mg bid) or omeprazole (40 mg bid) and amoxicillin (1 g bid) and clarithromycin (500 mg bid). This will change as the organism evolves. Development of a prophylactic/therapeutic vaccine is underway.

GASTRIC CANCER—The two major types of gastric cancer are adenocarcinoma and lymphoma, both of which are seen most commonly with *H pylori* infections.

Adenocarcinoma occurs almost exclusively in the presence of gastric atrophy caused by either environmental gastritis (mostly *H pylori*) or autoimmune gastritis (pernicious anemia). It is usually, clinically silent until well advanced at which time patients present with weight loss (96%), pain (70%), vomiting (50%), anorexia (25%), early satiety (10%), hematemesis (10%), or dysphagia. Diagnosis is made by endoscopy with biopsy. The treatment is surgical with a 5-year survival rate of only 5–10%.

Lymphoma is the second most common malignancy in the stomach. The stomach is ordinarily devoid of lymphatic tissue, thus, lymphomas comprise less than 5% of all gastric malignancies. Most lymphomas are derived from mucosa-associated lymphoid tissue (MALT) and are B cell tumors. More than 90% are associated with *H pylori*. They may be associated with abdominal discomfort, nausea, vomiting, weight loss, or hemorrhage. Low-grade tumors regress after *H pylori* eradication. More advanced tumors require surgical resection followed by combined radiation therapy and chemotherapy.

Pancreas

The pancreas is located in the retroperitoneal space at approximately the level of the 2nd and 3rd lumbar vertebrae. The head of the pancreas fits into the C-loop, ie, the second portion of the duodenum. The body extends across the spine behind the stomach, and the tail lies in the hilus of the spleen. The pancreas has both exocrine and endocrine function.

Normal Physiology—The endocrine functions of the pancreas are mediated by hormones secreted by the islets of Langerhans. These cells account for less than 1% of the total mass of the pancreas and are scattered erratically throughout the gland. Within the islets are four distinct types of cells. The beta cells comprise 80% of the islet cell mass and secrete insulin. Alpha cells are found in the periphery of the islets and make up 16% of its mass. They secrete glucagon. Delta cells secrete somatostatin and the newly recognized polypeptide cells secrete yet to be identified products.

Pancreatic exocrine function is mediated by bicarbonate and digestive enzymes secreted into the intestine. Bicarbonate is secreted by the intralobular ductal cells. It provides an appropriate pH environment for pancreatic enzymes and protects the duodenal mucosa from acid from the stomach. There are more than 15 digestive enzymes that have been identified to date. These are produced in the pancreatic acinar cells. The most important are lipase, which cleaves triglycerides to form fatty acids and monoglycerides; amylase, which is responsible for the digestion of complex carbohydrates; and trypsinogen, which activates various protease enzymes that break down complex proteins.

Water and bicarbonate secretion are mediated by secretin, a 27 amino acid peptide secreted by S cells in the upper small intestine. Secretin release is induced by acidification of the duodenum. Pancreatic enzyme release is mediated by cholecystokinin, a 33 amino acid polypeptide release from mucosal cells in the upper small intestine in response to amino acids and triglycerides. Other hormones also are thought to play a role in pancreatic secretion although their precise function is not understood.

DISEASES OF THE PANCREAS

ACUTE PANCREATITIS—Acute pancreatitis is an acute inflammation of the pancreas. Gallstones and alcohol are the most common causes. Hyperlipidemia is an important and increasingly recognized cause of acute pancreatitis. It usually is associated with lipoprotein lipase deficiency and causes the most severe form of acute pancreatitis. Triglyceride levels are generally over 1000 mg/L in these patients. Other causes include trauma, vasculitis, infections (mumps and Coxsackie virus are the most common), spider bites and drugs (azathioprine, steroids, and thiazides are the most common).

The most common symptoms of pancreatitis are pain, nausea, and vomiting. The pain is usually mid-epigastric and bores through to the back. Fever may be present.

The diagnosis of acute pancreatitis is based on the clinical presentation and supported by a marked elevation of serum amylase or lipase. The white count usually is elevated, and mild jaundice may be present. X-rays of the abdomen usually show a dilated loop of bowel (so-called sentinel loop) near the pancreas. CT scan shows swelling of the pancreas.

Treatment of acute pancreatitis is supportive. Intravenous fluids are required. Nasogastric suction may be necessary to decrease nausea and vomiting. Pain is alleviated with narcotics. When patients are infected, antibiotics are given.

CHRONIC PANCREATITIS—Chronic pancreatitis is a chronic, relapsing inflammation of the pancreas that is manifested by recurrent episodes of abdominal pain, steatorrhea, and diabetes. The most important cause is alcoholism. The disease may be insidious in onset and present only with its end-stage manifestations of steatorrhea and diabetes. Bulky, foul-smelling, light-colored stools characterize steatorrhea. Malnutrition ensues from fat malabsorption, negative nitrogen balance, and diabetes. Malnutrition may be associated with weakness, anorexia, and signs of specific nutritional deficiencies. These include pathological bone fractures from vitamin D deficiency, bruising, and bleeding from vitamin K deficiency, night blindness from vitamin A deficiency, and muscle wasting and edema from protein deficiency. Pain may be a prominent feature of the disease. The treatment of chronic pancreatitis is directed toward the prevention of malnutrition and, if present, the relief of pain. Nutrition is restored with the use of good diet and pancreatic replacement enzymes. Pain management is very difficult in these patients because many are addicted. Narcotics should be avoided. There is evidence that pancreatic enzyme replacement relieves pain in some patients.

PANCREATIC TUMORS—There are two major types of pancreatic tumors: adenocarcinomas arising from ductular epithelium and islet cell tumors arising from cells in the islets of Langerhans. Adenocarcinoma of the pancreas is usually insidious in onset with nonspecific symptoms such as weight loss, mild abdominal pain, and back pain. Jaundice due to obstruction of the common bile duct ultimately ensues. Occasionally, systemic manifestations such as migratory thrombophlebitis, erythema multiforme, thrombocytosis, and fever of unknown origin occur. Pancreatic adenocarcinoma is almost invariably incurable at the time of diagnosis.

Patients with islet cell tumors frequently exhibit symptoms and signs related to the tumor secretory products. For example, hyperinsulinemia may produce hyperphagia, weight gain, and mental changes. Hypergastrinemia may be associated with aggressive ulcer disease. These tumors are frequently difficult to locate, often eluding CT scan and angiography. They are diagnosed most commonly based on the clinical history and measurement of their secretory products.

CYSTIC FIBROSIS—Cystic fibrosis is an inherited, autosomal recessive disease seen in about 1 in 1500 to 2000 live births. Severe pulmonary disease predominates, but there are also gastrointestinal manifestations, particularly steatorrhea with malnutrition. (See Respiratory section.)

Colon

The colon, or large bowel, is a 3- to 4-foot long tubular organ. It extends around the periphery of the abdominal cavity. Its primary functions are the reabsorption of water and electrolytes and the storage of feces for evacuation at a convenient time.

Normal Physiology—Approximately 1500 to 2000 ml of liquid chyme reaches the ileocecal valve each day. This is the net volume following ingestion, absorption, and secretion from the upper GI tract. The intestinal bolus empties slowly through the ileocecal valve into the cecum. In the ascending and transverse colon, the ring-like contractions further delay the movement of chyme. Sodium, followed by water, is absorbed actively in this part of the bowel, transforming the chyme into a soft, fecal mass. In the transverse and descending colon, tonic contractions carry the globular mass downstream, often propelling it distances that reach 1/3 the length of the colon. These mass movements frequently occur as part of the gastrocolic reflex after eating. Defecation is initiated by distention of the rectum by the fecal mass. If the urge to defecate is resisted, the stimulus gradually diminishes, and sometimes constipation ensues. The colon's contribution to water balance in the intestine is relatively minor. Approximately 10 L of fluid enters the gut daily. This consists of oral intake of 2 L, saliva of 1 L, gastric juice of 2 L, bile of 1 L, pancreatic juice of 2 L, and jejunal secretions of 2 L. Of this amount, 8 to 9 L are reabsorbed in the small intestine. Another 1 to 2 L is reabsorbed in the colon, leaving 100 to 160 ml to be excreted daily as stool. It follows that the volume of the stool aids in defining the site of bowel dysfunction, which results in diarrhea.

Large-volume diarrhea, ie, greater than 1 L per day, is usually due to a disorder of the small intestine, whereas small-volume diarrhea, consisting of less than 1 L per day, is usually of colonic origin. Diarrhea and constipation are difficult to define because the frequency and volume of defecation varies greatly among individuals and in varying parts of the world depending on the diet. In general, normal bowel activity is defined as between three bowel movements per day and three bowel movements per week.

SYMPTOMS OF DYSFUNCTION

CONSTIPATION—Constipation generally denotes the infrequent or difficult evacuation of feces. It is a symptom of a problem rather than a medical disorder itself. By far the most common cause is irritable bowel syndrome, but it also occurs in association with hypothyroidism, hyperparathyroidism, hypercalcemic states, neurological disorders, and psychiatric disorders and in association with many drugs. Minor episodes of constipation may occur with changes in diet, particularly a decrease in fiber intake, and with alterations in daily routines such as travel and decreased physical activity. It also may occur in disorders of anal function that accompany neuromuscular disorders of the anal area. The law of Laplace ($t = P.r$) describes the important relationship between the tension in the muscle wall (t), the radius of the bowel lumen (r), and the pressure in the lumen (P). It forms the rationale for treatment of constipation with increased fiber. The important point is that increased muscle contraction, particularly in the colon, increases intraluminal pressure and retards the forward movement of feces, thus increasing the contact time for the reabsorption of water and the hardening of the stool. An increased fiber diet increases luminal diameter, thus decreasing intraluminal pressure and allowing more forward flow of the feces. Thus, fiber-containing laxatives form the most physiological basis for relieving constipation.

DIARRHEA—Diarrhea is defined as increased frequency or decreased consistency of bowel movements. It usually is classified as either of small bowel or large bowel origin. Small bowel diarrhea is usually large volume, consisting of large rushes and is associated with periumbilical cramping. Colonic diarrhea is small volume consisting of small spurts and associated with hypogastric cramping. Diarrhea is classified further as osmotic or secretory. Osmotic diarrhea is typically smaller volume, aggravated by eating and partially relieved by fasting. Secretory diarrhea is usually large volume and persists with fasting. It is possible to distinguish osmotic and secretory diarrhea by measuring stool osmolality. However, the logistics of such an examination make it difficult at best and almost routinely inaccurate in most clinical settings.

The major causes of osmotic diarrhea are inflammatory bowel disease, intestinal lactase deficiency, and various infections. The major causes of secretory diarrhea, which is uncommon, are villous adenoma and the various hormonal syndromes from non-GI tumors that secrete peptides that stimulate intestinal water secretion.

DISEASES OF THE COLON

IRRITABLE BOWEL SYNDROME—Irritable bowel syndrome is the most common chronic G-I disorder in the western world affecting close to 20% of those living in the US. It is characterized by intermittent abdominal pain, bloating, complaints of excess gas, food intolerance, and disordered bowel function consisting of either diarrhea or constipation or, most typically, both. The symptoms are thought to be the consequence of altered bowel motility, although specific disorders of motility have not been identified. The pain typically occurs in the lower

abdomen or the left- or right-upper quadrant. It is intermittent and often relieved by bowel movement or passage of flatus. It does not awaken the patient at night. When the pain occurs under the left coastal margin, it is known as splenic flexure syndrome, and when it occurs under the right costar mar gin, it is known as hepatic flexure syndrome. The diagnosis is made primarily on the basis of symptoms. It frequently occurs during stressful periods in people s lives or with changes in lifestyle with subsequent alterations in diet, particularly a change to diets that are low in fiber. It also is seen frequently with pharmacologic therapy, especially drugs with anticholinergic activity such as tricyclic antidepressants or major tranquilizers. Patients less than 30 years of age can be treated without diagnostic *workup*, but for those over 30, sigmoidoscopy and microscopic stool exam should be included. It is also important in these patients to rule out intestinal lactase deficiency.

The treatment of irritable bowel syndrome is reassurance, dietary modification to a regular high-fiber diet and fiber supplementation with bulk laxatives. Occasionally, antidepressants are needed for patients who are depressed. It is desirable to avoid antidepressants with anticholinergic activity in such patients.

DIVERTICULOSIS AND DIVERTICULITIS—Diverticula are acquired herniations of the mucosa through the muscular layers of the bowel. Diverticulitis is inflammation in a diverticulum resulting from microperforation. Diverticula may be the ultimate expression of irritable bowel syndrome and are rare before age 35 but present in 40–50% of people over 70. They are most common in the sigmoid colon, which has the highest intraluminal pressure. Diverticula are usually asymptomatic although they occasionally bleed. The treatment of diverticulosis consists of a high-fiber diet as used in the management of irritable bowel syndrome.

Diverticulitis, resulting from a perforation of a diverticulum, occurs in only 10–20% of people with diverticula. It manifests with acute, left lower-quadrant abdominal pain, fever, and constipation. Barium enema or colonoscopy usually makes diagnosis. The treatment of diverticulitis consists of the administration of antibiotics and, initially, a low residue diet consisting of enteral formulas. Once recovery occurs, the treatment is the same as that for diverticulosis.

ULCERATIVE COLITIS—Ulcerative colitis is a chronic disease of unknown etiology. It is an immune-mediated disease, but it is not known what triggers the immune response. The disease occurs predominantly in adults, 20 to 50 years of age, but may be seen at any age. It is more common in women, Caucasians, and Jews and in those who reside in urban settings. It is rare among Africans, Asians, and North American Indians

The pathology of the disease is very characteristic. The mucosa of the rectum and bowel is edematous with a bloody purulent exudate. The rectum is virtually always involved with the disease, having a tendency to spread from the rectum to more proximal areas in a continuous pattern.

Bloody diarrhea is the most characteristic presentation of ulcerative colitis. The stool also may be purulent. Diarrhea with as many as 20 to 30 bowel movements per day is common. Lower abdominal pain, hematochezia, and fever also occur. Laboratory data usually show leukocytosis and anemia. Diagnosis is made by sigmoidoscopy with mucosal biopsy.

The clinical course of ulcerative colitis is variable but intractable. Spontaneous remission does occur but, in general, the course of the disease is one of exacerbations and remissions. It is a lifetime disease. Because of the risk of colon cancer and the superimposition of complications, most patients have a total colectomy within the first 10 years of the onset of disease.

Complications include perforation with peritonitis, toxic megacolon resulting from a dilated functionless bowel, and adenocarcinoma of the colon. The risk of adenocarcinoma increases with age. It is about 2–3% at 10 years and 20–25% after 20 years of disease. The diagnosis of carcinoma in the presence of ulcerative colitis is difficult because the symptoms of ulcerative colitis mask the symptoms of carcinoma. Because of the difficulty in diagnosing colon cancer in patients with ulcerative colitis, the diagnosis often is delayed, and the mortality rate is greater than 50%.

Extracolonic manifestations also occur and include erythema nodosum, pyoderma gangrenosum, uveitis, iritis, and a variety of liver diseases including chronic hepatitis and primary sclerosing cholangitis, which, ultimately, usually requires liver transplant.

The treatment of ulcerative colitis consists of drugs that reduce the inflammation. These include corticosteroids, azathioprine, and methotrexate. In addition, intestinally acting preparations of salicylate such as azulfidine, olsalazine, and mesalamine are used. The latter drugs are particularly useful in reducing the frequency of flare-ups of disease. The ultimate treatment, however, is total colectomy with ileo-anal pull through.

CROHN'S DISEASE (GRANULOMATOUS COLITIS)—Crohn's disease is a granulomatous inflammation that affects both the colon and small bowel. When it involves only the colon, it is frequently indistinguishable from ulcerative colitis. Like ulcerative colitis, the etiology is unknown, but immune mechanisms appear to be important. The clinical and laboratory features of Crohn's colitis are indistinguishable from ulcerative colitis. Distinction is made by bowel biopsy, which may show the characteristic granulomatous inflammation. When that inflammation is not present, Crohn's may be indistinguishable from ulcerative colitis for several years. The complications of Crohn's colitis are the same as those for ulcerative colitis.

The medical treatment of Crohn's colitis is the same as that for ulcerative colitis. However, in Crohn's colitis, every effort is made to preserve the colon, since surgery has a tendency to chase the disease up the bowel. Surgery in Crohn's disease is indicated only for complications such as perforation and stricture.

POLYPOID LESIONS OF THE COLON—Colonic polyps are very common. The adenomatous polyp is the most important polyp, affecting more than 20% of the population. Because of the frequency with which it occurs and because it is the precursor of colon cancer, adenomatous polyps are the targets of colon cancer screening.

Colonic adenomatous polyps are seen in 5–15% of the general population over 45 years of age, and prevalence increases with age. Adenomas usually are found during screening examinations for colon cancer, but may also present with symptoms of rectal bleeding, abdominal pain or diarrhea. Most patients with polyps will have 1 to 3, but as many as 50 may be seen. Most are pedunculated and can be removed through the colonoscope by snare electrocautery. Following removal, they tend to recur, and thus, follow-up examinations are important.

COLON CANCER—Malignant lesions of the colon include adenocarcinoma, lymphoma, sarcoma, carcinoid tumors, and, rarely, metastatic tumors. However, 95% of colon malignancies are adenocarcinomas. There are approximately 150,000 new cases with 60,000 deaths per year in the US. It is the second most common cause of cancer death in men (following lung) and women (following breast). One in 20 Americans will develop this malignancy. The incidence of colon cancer increases with age and is most common in the seventh decade.

Both environmental and genetic factors have been implicated in the cause of colon cancer. A high incidence has been linked to low dietary fiber intake and high animal fat consumption. An increased prevalence of colon cancer in relatives of colon cancer patients indicates that genetic factors are also important.

Colon cancer may cause blood in the stools, a change in bowel habits, abdominal pain, and/or weight loss. In most patients, however, the symptoms are late. Thus, most cancers are not resectable for cure by the time they become symptomatic. Because of the frequency of colon cancer and its curability when detected early, routine screening is indicated. The current recommendations for screening include yearly rectal exam after age 40, stool Hemoccult testing yearly after age 50, and proctoscopic exam at age 50 and every 3 to 5 years thereafter. Widespread screening has been shown to reduce the death rate from colon cancer.

Liver

The liver is responsible for the synthesis of cholesterol, bile salts, phospholipids and various proteins. It also stores and transforms carbohydrates. A major function of the liver is the detoxification and excretion of exogenous substances.

Amino acids are synthesized by the liver to tissue and plasma proteins, especially albumin. It also synthesizes nonessential amino acids as well as all of the coagulation factors except Factor 8. Glucose is stored in the liver as glycogen. A visible function of the liver is its conjugation of bilirubin, a product of hemoglobin degradation. The liver converts bilirubin to a polar form that can be excreted in bile and to some extent in the urine. Failure to metabolize bilirubin results in jaundice, a yellow discoloration of the skin and sclera that is a common symptom of liver disease.

Virtually all lipid-soluble exogenous substances are metabolized in the liver. This function is carried out largely by hydroxylation by the mixed-function oxidases, followed by conjugation. This process is responsible for most drug metabolism and is at the heart of many drug interactions.

The liver is unique in having two blood supplies. The veins from the GI tract and spleen form the portal vein, which perfuses the liver and normally accounts for about 70% of its blood supply. The liver also receives arterial blood from the hepatic artery. Approximately one-fifth of cardiac output normally flows through the liver.

The liver has a limited number of ways of responding to injury. These include acute hepatitis, chronic hepatitis, and fibrosis and tumor formation. In addition, there are a number of storage diseases of the liver. The remarkable ability of the liver to regenerate spares it from end-organ failure in most of these diseases.

DISEASES OF THE LIVER

ACUTE HEPATITIS—Acute hepatitis is caused by either viruses or toxins. The most important causes of acute hepatitis in the US are hepatitis A, B, and C viruses. The important drugs causing hepatitis are halothane, isoniazid and acetaminophen. The various types of viral hepatitis are compared in Table 56-1.

HEPATITIS A—Hepatitis A virus was first identified in 1973. Type A hepatitis occurs in epidemics, particularly in younger people. There is, however, a disturbing trend toward an increased age of acquisition. This is particularly problematic because hepatitis A virus, while causing a mild, flu-like illness in children, causes a very serious illness in middle-aged and older adults.

In the typical clinical course, a prodrome of malaise, anorexia, headache, mild fever, and alteration of taste occurs 6 to 8 weeks after exposure. Soon after, the patient notices dark urine, light stool, and some right-upper quadrant discomfort. Jaundice may follow after a few days. It is noteworthy that only a small percentage of patients actually become jaundiced; thus, the illness tends to be missed and attributed to flu. While hepatitis A occasionally becomes fulminant and causes death, the overwhelming majority of patients, (ie, greater than 99%) recover without sequelae.

HEPATITIS B—The hepatitis B virus was discovered in the mid 1970s. Hepatitis B virus is a DNA virus that has a tendency to cause chronic disease. The acute illness is indistinguishable from other types of viral hepatitis, but about 10% of patients develop a chronic hepatitis with some of these going on to ultimately develop cirrhosis or hepatocellular carcinoma. The primary mode of transmission throughout the world is vertical (ie, from infected mother to newborn infant), but in the US, the primary modes of transmission are sexual and IV drug abuse.

There is now an effective vaccine for hepatitis B. At present, only people at high risk of acquiring the disease are being vaccinated, but it is hoped that universal vaccination will be underway soon.

HEPATITIS C—Hepatitis C virus is the major cause of transfusion-associated hepatitis, although transfusion is not the major mode of spread. Over 50% of cases are acquired by an unknown mode of transmission. The clinical course of hepatitis C virus is indistinguishable from other forms of viral hepatitis, although it tends to be milder. The most characteristic feature of this illness is its propensity to become chronic. At least 70% of patients who are infected, ultimately developing chronic disease, and about 20% of these ultimately going on to develop liver failure or liver cancer.

There is no vaccine to prevent hepatitis C. There is some evidence, however, that treating the acute illness with alpha-interferon reduces the incidence of chronic disease.

PREVENTION OF VIRAL HEPATITIS

Hepatitis A—Optimum control lies in good general hygiene, safe disposal of feces, and identification of epidemics. Immune serum globulin is effective in preventing or modifying type A hepatitis in over 50% of those exposed. A worrisome feature, however, is that with the declining incidence of hepatitis A in the young population, less and less of the pooled immune specific globulin is effective in preventing hepatitis A. A hepatitis A vaccine has been developed and should be available commercially in the near future.

Table 56-1. Comparison of Types of Hepatitis

FEATURE	A	B	C	D	E
Virus	RNA	DNA	RNA	RNA	RNA
Incubation					
Range (days)	15–50	30–150	15–160	30–150	20–40
Mean (days)	30	75	50		27
Transmission					
Fecal-oral	Yes	No	Min[a]	?	Yes
Household	Yes	Min[a]	Min[a]	?	Yes
Vertical	No	Yes	? Min[a]	?	?
Blood	Rare	Yes	Yes	Yes	No
Sexual	No	Yes	Min[a]	?	?
Carrier state	No	Yes	Yes	Yes	No
Risk of chronic hepatitis	No	10%	70%–90%	Yes	No
Risk of liver cancer	No	Yes	Yes	?	No
Prevention					
Vaccine	Yes	Yes	No	No	No
Immunoglobulin	Yes	Yes	No	No	?
Mortality rate	≈0.15%	≈1%	≈0.5%	High	0.5%–1.5%

[a] Min = minimal.

Hepatitis B—Avoidance of multiple sexual partners and IV drug use is the most useful way of preventing hepatitis B. Hepatitis B immune specific globulin (HBIG) appears to be effective in preventing hepatitis B in about 75% of cases. There is also an effective vaccine for preventing hepatitis B (Energix B or Recombivax BB).

Hepatitis C—There is no known mechanism for preventing hepatitis C. Pooled immune globulin is not effective. There is as yet no vaccine.

Chronic Hepatitis—Chronic hepatitis is the pathological and clinical manifestation of a heterogeneous group of disorders, both genetic and acquired. What they have in common is a chronic inflammatory reaction directed against the hepatocyte. By far the most common causes are hepatitis B and C viruses, which account for 70–80% of cases in most series. Autoimmune chronic hepatitis, Wilson's disease, and drugs account for the remainder. The disorders can be distinguished on the basis of several serological tests. Our understanding of these diseases has evolved largely over the past 20 years and was propelled by the discovery of the hepatitis viruses.

Chronic infection with hepatitis B is the most important worldwide cause of chronic hepatitis. The liver injury results from an inflammatory immune attack against hepatocytes. In most patients, the hepatitis B virus itself is not cytopathic. The infected cells are not eliminated, allowing the attack to continue. In the usual circumstance, the hepatocyte expresses cell surface markers (in this case HBcAg and HLA Class I antigen). Primed lymphocytes then attack the infected hepatocytes. The expression of the HLA markers is stimulated by interferon. There is now considerable evidence that patients with chronic hepatitis B are deficient in interferon and, by inference, unable to express HLA markers that would attract an appropriate lymphocyte response. This deficiency is probably genetic in some populations and acquired in others. The acquired deficiency occurs as a consequence of transfection of chromosome 9 at the site that codes for interferon.

The discovery of interferon deficiency in chronic hepatitis B has led to the successful use of interferon as therapy in some of these patients. Approximately half of the patients respond with a loss of viral replication, a reduction in inflammation and, in some cases, a loss of the markers of hepatitis B infection including HBsAg. In general, patients with aminotransferase enzyme (ALT or AST) levels of 100 to 200, DNA levels of less than 100 and positive HBeAg respond best. The treatment is 5 million units subcutaneously daily for 6 months. At about the 12th or 14th week, one can expect to see a flare-up of the hepatitis. This is a good sign and usually associated with conversion of HBeAg to anti-HBe and loss of viral replication. The response, when obtained, usually is prolonged with a relapse rate of only 2–3% per year.

WILSON'S DISEASE—Wilson's disease is an autosomal recessive disorder of copper metabolism that manifests primarily as either neuropsychiatric disease or liver disease. It has a gene frequency of 1/200 and a disease frequency of 1/30,000. More than 30 different mutations on chromosome 13 have been found. Wilson's disease usually presents prior to age 30, although several patients in their 50s and 60s have been reported. For reasons that are unknown, children tend to have predominantly hepatic involvement while adolescents and adults have the neuropsychiatric manifestations. The hepatic manifestations include fulminant hepatitis, chronic hepatitis, and cirrhosis. Hepatocellular carcinoma is virtually unknown in Wilson's patients. Approximately 25% of patients have evidence of involvement of more than one organ system at the time of diagnosis. The characteristic laboratory features include moderately elevated aminotransferase enzymes (2–5 fold), normal or near normal alkaline phosphatase and absence or near absence of the copper carrier protein, ceruloplasmin. The role of ceruloplasmin in the pathogenesis of Wilson's disease is unknown. The ceruloplasmin gene, however, is on chromosome 3, rather than 13, and thus the deficiency of ceruloplasmin is probably a secondary feature. The underlying pathophysiology, whatever the mechanism, is an inability to excrete biliary copper that accumulates in various tissues leading to the characteristic clinical features consisting of neuropsychiatric changes including behavioral change, psychosis, extrapyramidal signs, and cerebellar or pseudobulbar signs. Corneal rings known as Kayser-Fleischer rings are virtually pathognomonic. However, they are frequently not present in younger patients with liver disease. Other manifestations of Wilson's disease include proximal renal tubular dysfunction, osteopenia, osteoarthropathy, and hemolysis.

The diagnosis is based on finding disturbances in copper metabolism including decreased or absent serum ceruloplasmin, urinary copper excretion of greater than 100 mg per day and hepatic copper concentration of greater than 250 μg/g of liver tissue.

Untreated Wilson's disease is fatal. The treatment consists of chelation therapy with D-penicillamine and is lifelong. Patients who develop fulminant hepatitis die unless they receive a liver transplant. Patients with chronic hepatitis eventually progress to cirrhosis despite treatment and eventually require liver transplantation.

AUTOIMMUNE CHRONIC HEPATITIS—This is also a heterogeneous group of disorders that can be distinguished on the basis of serological tests. It is not yet known, however, whether the different types of autoimmune hepatitis have different courses or response to treatment. It is less common than chronic hepatitis B or C. The typical clinical features are female predominance, young age, association with autoantibodies and other autoimmune disorders, presence of hyperglobulinemia, and virtually universal response to corticosteroids. It is associated with HLA phenotypes B8 and DR3. Interestingly, patients with either autoimmune or viral chronic hepatitis that is associated with other autoimmune disorders are more likely to be DR4 phenotype. The disease usually is progressive with development of cirrhosis and liver failure within a few years. Corticosteroids greatly improve the prognosis of autoimmune chronic hepatitis. The initial steroid therapy is titered to the serum aminotransferase enzyme levels. Patients should be maintained on low-dose steroids indefinitely after the initial response. Azathioprine may be used for a steroid-sparing effect.

Wilson's disease is a rare cause of chronic hepatitis. It usually occurs before age 30, but several patients in their 50s and 60s have been reported. For reasons that are not known, patients have predominantly either the liver or the neuropsychiatric form of the disease. In children, hepatic involvement tends to dominate, while in adolescents and adults the neuropsychiatric disease tends to dominate. Approximately 25% of patients have evidence of involvement of more than one organ system at the time of diagnosis. The consequence of missing the diagnosis is disastrous with virtually all patients subsequently developing acute liver failure. Patients with Wilson's disease tend to have normal or near-normal serum alkaline phosphatase and alanine (ALA) levels. They also tend to have periportal Mallory's hyaline on liver biopsy, unlike other forms of chronic active hepatitis (CAl1). Early intervention with chelation therapy (D-penicillamine) leads to stabilization and improvement in the liver disease. Development of fulminant hepatic failure is always fatal and an indication for emergency liver transplantation.

The final cause of chronic hepatitis among the major categories is drug-induced. A number of drugs have been reported including methyldopa, nitrofurantoin, isoniazid, ketoconazole, and acetaminophen Women appear to be more susceptible, and there is frequently a background of autoimmune disease. The clinical presentation mimics autoimmune chronic hepatitis. The treatment is drug withdrawal.

CIRRHOSIS—Cirrhosis (Gk, kirrhos = yellow) is defined as a diffuse increase in fibrous tissue within the liver plus the presence of regenerative nodules. The fibrosis is the result of active fibrogenesis. The fibrogenesis generally is thought to be stimulated by cytokines released during inflammation and necrosis. Virtually all chronic liver diseases ultimately can end with cirrhosis. The fibrous tissue leads to a distortion of the architecture of the liver with loss of normal function. Even though regeneration of hepatocytes occurs, the distorted architecture compromises their overall function.

By far the most common cause of cirrhosis in this country is alcohol consumption. Other causes include chronic active hepatitis of all types, primary biliary cirrhosis, hemochromatosis, Wilson's disease, and alpha-1 antitrypsin deficiency. The typical patient with alcoholic cirrhosis has consumed approximately a pint of whiskey per day for 15 years. However, the majority of patients who drink this much alcohol never develop cirrhosis. It probably is determined genetically whether or not

cirrhosis occurs. In the case of alcoholic cirrhosis, only about 20% of patients who are alcoholic develop cirrhosis.

The clinical presentation of cirrhosis is related primarily to the development of portal hypertension and the loss of hepatocellular function.

Portal hypertension results from the resistance of flow through the liver. The increased pressure in the portal system is transmitted within that system, especially the coronary vein leading to esophageal varices, the gastric veins leading to gastric varices, and the inferior mesenteric vein leading to hemorrhoids. When the pressure reaches a certain level, these veins tend to burst, causing gastrointestinal hemorrhage. This is particularly true for the esophageal varices.

Another manifestation of cirrhosis is ascites, the accumulation of fluid in the abdominal cavity. The pathophysiology of ascites formation is complex, but the two most important features appear to be an increase in hydrostatic pressure in the portal circulation as the consequence of portal hypertension and decreased oncotic pressure due to the development of hypoalbuminemia. The hypoalbuminemia is caused by decreased synthesis of albumin by hepatocytes and the loss of albumin from the surface of the liver. This results in decreased oncotic pressure in the circulation (from decreased albumin synthesis) and increased oncotic pressure in the free peritoneal space (from albumin in the perinatal space). These factors in combination favor fluid accumulation in the abdominal space. The loss of fluid from the intravascular space causes secondary hyperaldosteronism, which activates the renin angiotensin system causing the kidneys to retain sodium and water. Thus, a vicious cycle is formed, all directed toward fluid retention.

Porto-systemic encephalopathy (PSE) is another manifestation of cirrhosis and is characterized by a spectrum of decreased mental and neurologic function. PSE is thought to occur because of the failure of the liver to remove noxious products of protein metabolism, particularly ammonia. Typical symptoms include sleep reversal, hypersomnia, apathy, personality changes, and intellectual deterioration. There may be neurological abnormalities such as slurred speech, asterixis, and exaggerated deep-tendon reflexes. The diagnosis is made on the basis of the clinical presentation and a characteristic delta wave pattern on electroencephalogram.

Other clinica1 features include the manifestations of excess feminization due to the toxic effect of alcohol on testicular function and the failure of the liver to metabolize estrogen. The net effect of excess feminization is spider angioma, palmar erythema, Dupuytren's contracture, parotid enlargement, gynecomastia, and testicular atrophy.

Symptoms and Signs—The most characteristic manifestations of cirrhosis are jaundice and ascites. However, an insidious onset characterized by weakness, fatigue, anorexia, and ultimately the signs of PSE, including sleep reversal, apathy, forgetfulness, confusion, euphoria, and personality changes, may occur. Social graces are often lost. Stupor and coma eventually ensue. Neurological findings, at this time, might include asterixis, slurred speech, muscle rigidity, hyperreflexia, and occasionally, localizing neurological signs. Primary biliary cirrhosis may have some unique features such as pruritus, dark urine, pale stools, steatorrhea, and xanthelasma.

Laboratory abnormalities include hyperbilirubinemia, hypoalbuminemia, prolonged prothrombin time, hyponatremia, and mildly elevated AST and ALA levels. Pancytopenia may be present. In primary biliary cirrhosis, the serum alkaline phosphatase is elevated markedly as is the serum cholesterol. Antimitochondrial antibodies are present in the serum.

The clinical course of cirrhosis is usually relentlessly downward. In alcoholic patients, this downward course may continue despite abstinence. The fatal event is usually bleeding from esophageal varices or an infection.

There is no specific curative treatment for any form of cirrhosis. However, the prognosis in alcoholic cirrhosis is improved by abstinence. The prognosis in autoimmune chronic active hepatitis is improved by continuous low-dose corticosteroid therapy. A preliminary study has shown methotrexate to be partially effective in the treatment of primary biliary cirrhosis. The cirrhosis of hemochromatosis is treated by iron removal by phlebotomy, but there is little evidence that once the patient has become cirrhotic that the prognosis is improved. The prognosis of Wilson's disease is improved with copper chelation therapy with D-penicillamine. Preliminary studies have shown that the course of chronic hepatitis B may be improved with alpha-interferon therapy. Nevertheless, liver transplantation remains the treatment of choice for patients with end-stage liver disease.

Gallbladder and Gallstones

The gallbladder stores and concentrates bile. It is the usual site of gallstone formation.

Normal Physiology—The gallbladder fills passively with bile secreted by the liver. The filling process is facilitated by the secretion of bile and the closing of the sphincter of Oddi between meals, which enables the gallbladder to fill with bile, concentrate the bile, and then contract after meals to empty into the intestine where the bile solubilizes lipids for ultimate digestion and absorption. The gallbladder contracts and empties its concentrated bile in response to cholecystokinin released from the duodenal mucosa during a meal.

Bile is the major secretory product of the liver. It is composed of water in which small amounts of cholesterol, phospholipids, and bile salts are solubilized. It also contains bilirubin, which gives bile its characteristic yellow color. Bile is increasingly concentrated as it proceeds through the biliary tree and is concentrated 10- to 20-fold in the gallbladder, which absorbs water. Cholesterol is insoluble in water but is dissolved in bile by incorporation into mixed micelles and small vesicles. Mixed micelles are composed of bile acids, which are detergents, and lecithin, which together solubilize cholesterol. There is a limit to the quantity of cholesterol that can be dissolved in micelles. If this quantity is exceeded, cholesterol precipitates, which predisposes to gallstone formation.

Bile acids are synthesized from cholesterol in liver cells. The primary bile acids, cholic acid and chenodeoxycholic acid, are conjugated in the liver, excreted into the bile, and eventually reach the small intestine, where they participate in the solubilization of lipids. About one-third of the primary bile acids secreted into bile are converted by intestinal bacteria to the secondary bile acids, lithocholic acid and deoxycholic acid, which are lost in the stool. The remaining primary bile acids are reabsorbed in the terminal ileum and returned to the liver to be recycled—the enterohepatic circulation. This mass of recirculating bile acids, called the bile acid pool, recirculates approximately twice with each meal. Most of the reabsorption takes place in the last 100 cm of the terminal ileum, leaving a high concentration of bile acids to participate in digestion in the jejunum and proximal ileum. Loss of the last 100 cm of the terminal ileum, as occurs with surgery or regional enteritis (Crohn's disease), leads to malabsorption of fats, decreased fat absorption, and diarrhea (induced by bile acids in the colon).

CHOLELITHIASIS (GALLSTONES)—Gallstones are classified according to their composition: cholesterol, pigment, and mixed. Mixed stones are by far the most common. They are predominantly cholesterol but also contain bile pigments, calcium salts, and protein. They probably have a pathogenesis similar to that of pure cholesterol stones. They are often multiple, with a brown center, hard shell, and faceted surface. Pigment stones contain bile pigment such as bilirubinate. They are black, round to amorphous, and hard. Two-thirds of gallstones in the US are predominantly cholesterol.

Epidemiology—An estimated 24 million Americans have gallstones. In those over age 65, the incidence approaches 30%. Cholesterol and mixed stones are three times more common in women of child-bearing age than in men. The incidence is increased in individuals who are obese, elderly, multiparous, or cirrhotic. The incidence exceeds 70% in women of some Native American tribes.

Pathophysiology—The pathogenesis of cholesterol gallstone formation has been clarified. Failure of cholesterol-volatilization leads to precipitation and potentially to a gallstone. Normal people may secrete iatrogenic bile (supersaturated with cholesterol) during fasting when bile acid secretion is minimal but not all people develop gallstones. Nevertheless, certain defects have been identified in patients with cholesterol gallstones. Lean people with gallstones tend to have reduced biliary secretion of bile acids and phospholipids. Obese individuals secrete excessive quantities of cholesterol into bile. Some individuals have a contracted bile acid pool because their bile acid loss exceeds the maximum rate of liver synthesis of bile acids. For example resection or chronic inflammatory disease of the ileum may cause the net loss of bile acids as may the chronic ingestion of the binding resin, cholestyramine.

Once a crystal is formed as a result of cholesterol precipitation from bile, the crystal may grow or several crystals may aggregate. This phase of gallstone formation is poorly understood. Nucleating factors exist in bile and appear to foster precipitation of cholesterol crystals. The process of gallstone growth appears to involve the entrapment of crystals

by gallbladder mucus, and the process may be fostered by impaired gallbladder emptying.

Information regarding pigment stone formation is scarce. Many patients have increased bilirubin production as a result of chronic hemolysis. Thus, the liver conjugates and excretes increased quantities of bilirubin. Beta-glucuronidase in bile may deconjugate bilirubin, making it less soluble in bile and possibly fostering precipitation.

Gallstones cause morbidity by irritating the gallbladder mucosa directly (cholecystitis) or by impacting in the cystic duct. They also may pass into and obstruct the common duct.

Symptoms and Signs—Most patients with gallstones are asymptomatic. The characteristic symptom is epigastric pain that may lateralize to the right side and radiate to the tip of the right scapula. The pain is a severe, aching sensation that is not influenced by body position. The pain begins rapidly, grows in intensity, and disappears rather abruptly. The duration of pain is variable but usually is about 2 to 6 hours. Nausea and vomiting may accompany the pain. Jaundice may appear in several days if the stones remain in the common bile duct. Fever and chills often occur with acute cholelithiasis because of infection in the biliary tree. Sepsis may occur. The symptoms of flatulence, bloating, and fatty food intolerance, frequently attributed to gallbladder disease, are not characteristic of gallbladder disease and are more likely due to irritable bowel syndrome.

Physical examination in the acute case reveals tenderness, muscle guarding, and rigidity over the area of the gallbladder. A mass is rarely palpable. Serum levels of alkaline phosphatase and bilirubin may be increased; WBC count is elevated in infection. Ultrasound discloses gallstones in most cases.

RENAL DISEASE

Normal Physiology—The kidneys receive about 20% of the resting cardiac output. From this torrential blood flow, the one million glomeruli in each kidney create an ultrafiltrate (glomerular filtrate) at a rate of 120 ml/minute. The glomerular filtrate contains all small molecules in the same concentration as they are dissolved in the plasma but does not allow the escape of large molecules (protein). Each glomerulus is connected to a renal tubule. The tubule reabsorbs about 99% of the glomerular filtrate and most of the dissolved solutes, returning to the bloodstream what is required for maintenance of the internal environment and allowing any excess to escape into the urine. Waste products such as urea and creatinine are reabsorbed to a much lesser extent or not at all and are thus preferentially eliminated. Relevant details of physiology are included in the appropriate sections below.

Glomerular Disease

As might be expected from the physiology above, disease of the glomeruli tends to reduce glomerular filtration rate and to allow leakage of protein into the urine. Common features of glomerular disease thus include fluid overload, hypertension, proteinuria, and renal failure. Diabetes is the most common cause of glomerular disease in western countries. Most other forms of glomerulonephritis (GN) involve immunologically mediated inflammation of the glomeruli in both kidneys symmetrically. GN must be differentiated from interstitial nephritis, which is inflammation of the connective tissue surrounding the glomeruli and tubules.

DIABETIC NEPHROPATHY

Definition/Overview—Diabetic nephropathy is characterized clinically by a stereotyped march from normality to subtle increase in glomerular filtration rate (hyperfiltration) to excretion of albumin in minimally increased quantities (microalbuminuria) to heavy urinary protein loss, and eventually decline of renal function to uremia.

Epidemiology—In the absence of effective treatment, 25–45% of patients with type 1 or type 2 diabetes will develop nephropathy during their lifetime. Certain groups (Native Americans) not only have a higher prevalence of diabetes than average, but also a greater likelihood of developing nephropathy

Pathology and Pathogenesis—Glycosylation of tissue proteins appears to lie at the root of the microvascular damage in diabetic nephropathy. High blood pressure and activation of cytokines including TGF-β magnifies the damage. By electron microscopy, uniform thickening of the glomerular basement membrane, diffuse expansion of the mesangium, and later the appearance of glomerular nodules (Kimmelstiel-Wilson lesion) characterize diabetic nephropathy.

Symptoms and Signs—Typically symptoms and signs develop late in the evolution of the renal disease. Hypertension, proteinuria, nephrotic syndrome, and chronic renal failure (see descriptions below).

GLOMERULONEPHRITIS

Etiology—Glomerulonephritis has diverse causes. Several potential immunological mechanisms can give rise to glomerulonephritis. For example, in Goodpasture syndrome or antiglomerular basement membrane nephritis, an endogenous antigen attaches to the basement membrane of glomerular capillaries and incites a destructive inflammatory nephritis. In lupus nephritis and postinfectious glomerulonephritis, antigen-antibody complexes deposit and initiate inflammation. In lupus nephritis, the antigen is DNA; in postinfectious glomerulonephritis, the antigen is a protein associated with the organism infecting some other part of the body; streptococcal antigen is a well-researched example. In IgA nephropathy, the immune mechanism is not clear. Broadly similar glomerular damage can also occur from non-immunological mechanisms, for example in vasculitis and hereditary nephritis (Alport syndrome).

Epidemiology—Glomerulonephritis is the leading cause of chronic renal failure after diabetes. IgA nephropathy is the most common cause of glomerulonephritis worldwide and is particularly common in Asians. Glomerulonephritis occurs in two-thirds of patients with lupus.

Pathology—Diverse types of histological damage reflect the diverse etiologies. In acute GN, such as poststreptococcal, the glomeruli are swollen, infiltrated with PMNs, and there is proliferation of endothelial and epithelial glomerular cells. In severe cases, epithelial crescents form in Bowman's capsule. In immune complex GN granular, nodular, or "lumpy bumpy" deposits of immunoglobulin are found in the glomeruli. In antiglomerular basement membrane nephritis, immunofluorescence microscopy shows antibodies in a linear pattern along the capillary walls of the glomeruli.

The pathological classification of chronic GN includes IgA nephropathy, membranoproliferative, membranous, focal or diffuse proliferative and rapidly progressive GN. A description of the histopathological features of these forms of GN is beyond the scope of this chapter.

Pathophysiology—All cases of GN are the result of immune reactions. Many cases involve formation of antibodies against circulating extrarenal antigens. These antibodies are usually IgG and also circulate in the blood. Antigen-antibody complexes are formed when a critical ratio of antibody to antigen is reached in the blood. The complexes become trapped in the glomeruli during filtration, hence the name immune complex glomerulonephritis. The process actually is more complex than simple trapping and involves dysfunction of the mesangial cells, the reticuloendothelial cells in the glomeruli that normally remove foreign materials. The antigen-antibody complexes in the glomeruli activate the complement cascade via the classic or alternate pathways. Activation of complement also activates Factor XII and the clotting system, which leads to the deposition of fibrin. Factor XII also activates the kinin system, which causes release of chemotactic factors, and substances that increase permeability of blood vessels. The inflammatory reaction with the release of lysosomal enzymes damages the glomeruli. Fibrosis ensues.

The remaining 5% of cases of GN are due to the development of antibodies against glomerular basement membrane. These antibodies also are active against alveolar basement membrane. The inflammatory reaction is responsible for the damage to the glomeruli and alveoli.

Symptoms and Signs—The hallmarks of GN are gross or microscopic hematuria (RBCs in the urine), hypertension, proteinuria, and facial, periorbital, and pedal edema. Edema is also part of the nephrotic syndrome and will be discussed below. Glomerulonephritis also may be associated with hypertension, fatigue, anorexia, and congestive symptoms such as orthopnea and dyspnea on exertion. The urine also may contain RBC casts, WBCs, granular or hyaline casts, and epithelial debris. Chronic GN eventually leads to the symptoms and signs of chronic renal failure.

Oliguria, "coke-colored" or "smoky" urine, bilateral steady flank pain, and malaise typically herald the onset of acute poststreptococcal glomerulonephritis. Edema develops in a few days unless salt and fluid are restricted.

The prognosis of acute poststreptococcal GN is excellent in children: 90% recover completely, although the urinary signs may persist for 1 year. The prognosis for chronic GN is variable. Some forms progress slowly while others deteriorate rapidly to chronic renal failure.

Nephrotic Syndrome

This is not a single disease but a constellation of abnormalities that occur when the glomerular capillary wall becomes permeable to protein.

Normal Physiology—Only small quantities of protein are filtered by normal glomeruli, a situation largely explained by the barriers to protein filtration and the nature of plasma proteins. The normal glomerular capillary wall is almost impermeable to protein. The endothelium is not a barrier, but the glomerular basement membrane prevents filtration of large proteins and blood cells. The negative charge on the glomerular basement membranes repels protein molecules. Thus, only proteins with a molecular weight of less than 40,000 may be filtered normally by the glomeruli, and the tubules reabsorb these proteins so that insignificant quantities of protein appear in the urine.

Etiology—Any glomerular disease that damages the basement membrane and allows leakage of protein may cause the syndrome. The most common cause of nephrotic syndrome in children is minimal change disease. In adults diabetes mellitus is far and away the most common cause; other causes include glomerulonephritis, amyloidosis, collagen vascular diseases, and nephrotoxins such as mercury, gold, anticonvulsant drugs, and penicillamine. Tubular disorders may cause mild to moderate proteinuria but do not cause nephrotic syndrome.

Pathophysiology—Large quantities of protein, mainly albumin, are lost in the urine in nephrotic syndrome. In adults the proteinuria is at least 3 to 4 g/day but may be as high as 30 to 40 g/day. Some filtered protein is degraded by tubules. Thus, measured proteinuria underestimates the total protein loss. Albumin synthesis by the liver can contend with a 15 g/day loss if dietary protein intake is adequate. When the loss exceeds the synthetic capacity of the liver, hypoalbuminemia occurs. Hypoalbuminemia results in a decreased oncotic pressure within blood vessels. Decreased oncotic pressure causes a decrease in fluid reabsorption in the venous capillaries resulting in edema. Loss of vascular fluid volume causes hypotension. The kidneys respond to the fall in blood pressure and volume by retaining sodium and water via the renin-angiotensin system. Up to 20 L of water may be retained in a futile attempt to restore blood volume, as the retained water simply becomes more edema fluid. Proteinuria leads to cast formation in the tubules. These may be hyaline, granular or waxy.

Hypercholesterolemia and hypercoagulability arise from overproduction of apolipoproteins and coagulation factors respectively Lipiduria also occurs, but not as a consequence of hyperlipidemia.

Symptoms and Signs—The classical symptoms and signs are proteinuria (greater than 3.5 g/m^2/day), hypoalbuminemia, and edema. The edema may be dependent and occur in the feet and ankles, or accumulate in compliant periorbital and facial tissue. The edema occasionally involves the entire body, a condition known as anasarca. Hyperlipidemia and lipiduria may or may not be present and are not essential for the diagnosis. Complications of nephrotic syndrome include hypotension and possibly shock, intravascular fluid overload or depletion, protein malnutrition, and a predisposition to thrombosis.

The prognosis is related to the prognosis of the underlying cause. However, the syndrome due to any cause may be fatal if fluid overload is not corrected.

Renal Failure

Renal failure is the inability of the kidney to perform its usual physiological functions and maintain homeostasis.

Renal failure may be classified as acute, subacute or chronic, depending on the time course.

Normal Physiology—The kidneys perform many functions. The fluid volume and serum osmolality are maintained by regulation of both sodium and water excretion. The pH of body fluids is maintained within narrow limits, normally 7.40 ± 0.04. The kidneys excrete many waste products.

The normal glomerular filtration rate (GFR) is 125 ml/minute and decreases with increasing age. The kidneys have a remarkable ability to adjust their excretion of water and solutes. They can excrete 20% of the glomerular filtrate if blood volume is expanded, which means that water intake could be as high as 35 L/day. The daily osmolar load obligates a urine output of 400 to 500 ml. The kidneys can excrete as much as 500 mEq of sodium/day or maintain sodium balance if intake of sodium is limited to 5 mEq/day. The kidneys normally excrete 50 to 80 mEq of potassium/day. The kidneys cannot produce urine virtually free of potassium, as they can in the case of sodium. This usually poses no problem as any mixed diet contains potassium, but it may prevent the kidneys from correcting hypokalemia if there are ongoing losses from the GI tract.

A person ingesting 70 g of protein forms 40 to 60 mEq of acid/day. The range of blood pH compatible with life is 6.9 to 7.6 but the normal range is much narrower. One half of the acid is excreted as titratable acid: $HPO_4^{2-} + H^+ \rightarrow H_2PO_4^-$. The other half is excreted by ammonia formation: $H^+ + NH_3 \rightarrow NH_4^+$. Filtered bicarbonate is reabsorbed completely unless the patient is alkalemic. The kidneys are responsible for excreting other waste products. Approximately 20% of filtered phosphate is excreted in the urine. A diet of 80 g of protein/day results in the formation of 20 grams of urea, which is excreted. The blood level of urea (blood urea nitrogen, BUN) is normally maintained below 20 mg/dl. The kidneys also excrete uric acid, magnesium, calcium, and other substances to maintain homeostasis.

The kidneys have several endocrine or metabolic functions. They produce erythropoietin, which regulates the red-blood-cell mass and renin, which regulates blood pressure and sodium and water balance. The kidneys degrade insulin and gastrin. The kidneys also participate in vitamin D metabolism and thus calcium homeostasis by converting a derivative of vitamin D, 25-hydroxycholecalciferol, to the biologically active form, I,25-dihydroxycholecalciferol.

ACUTE RENAL FAILURE—This is most commonly due to acute tubular necrosis (ATN) but also may be due to hypovolemia (prerenal azotemia) or to obstruction of the ureters, bladder or urethra. All excretory renal function can be lost within a few days.

Etiology—ATN is due most commonly to ischemia or toxins. Any event that leads to shock and intense vasoconstriction within the renal vascular bed may lead to it. Hemorrhage, hypotension during anesthesia, burns, sepsis, crush injuries, massive intravascular hemolysis, heart surgery requiring extracorporeal oxygenation, and childbirth may cause it. Toxins that may cause ATN include aminoglycoside antibiotics, radiographic contrast media, bichloride of mercury, carbon tetrachloride, ethylene glycol, methanol, myoglobin from crush injuries, and hemoglobin from intravascular hemolysis. Some cases have no identifiable cause.

Pathology—Ischemia causes patchy necrosis of the tubular epithelial cells and basement membrane. Other areas of the tubule may appear normal. Toxins cause diffuse necrosis of the tubular endothelial cells but do not injure the basement membrane. The glomeruli are spared in ATN unless the injury is severe and prolonged. The lesions are reversible if the patient survives.

Pathophysiology—Immediately after the injury renal blood flow may be reduced by as much as 50% by arteriolar constriction. Fluid filtered by the glomeruli leaks back into the interstitium through damaged tubules. The subsequent edema of the interstitium increases interstitial hydrostatic pressure, which further decreases renal blood flow and causes the tubules to collapse. Casts of degenerating epithelial cells block urine flow in the lumens and cause further increases in interstitial fluid. The kidneys can no longer maintain homeostasis by the excretion of sodium, water, and waste products.

Symptoms and Signs—Oliguria (urine volume of less than 400 to 500 ml/day) usually is the first sign of ATN but may not appear until several days after the injury. The composition of the urine formed is little changed from glomerular filtrate, but also contains protein and RBCs. The sodium concentration of the urine is fixed at about 50 mEq/L. BUN begins to rise and acidemia and hyperkalemia develop. If fluid therapy is not managed appropriately, hyponatremia and edema develop. The patient complains of nausea and lethargy. Death may occur within a few days because of acidosis and/or hyperkalemia.

During the second week, nausea, somnolence, weakness and thirst ensue. The BUN continues to rise and acidosis, edema, hyponatremia, and hyperkalemia worsen. Complications are common during this phase. Pulmonary edema, congestive heart failure, and hypertension may develop because of fluid overload. Hyperkalemia may cause cardiac arrhythmias. Metabolic encephalopathy, possibly due to urea, hyponatremia, and hypocalcemia results in neurological deterioration, convulsions, and coma. Anemia due to decreased RBC production, increased RBC destruction and dilution appears in the second week. Nosocomial infection is the most common cause of death in this phase.

During the recovery phase, urine volume increases daily. The BUN may continue to rise until urine volume has exceeded 1000 ml/day for several days. Polyuria (urine volume of greater than 3000 ml/day) may develop. Weight loss is rapid as the edema resolves. Since the tubules may not yet be able to conserve water, sodium or potassium, dehydration, hyponatremia and hypokalemia may develop. The diuresis may continue for 1 to 3 weeks. The GFR may never return to normal, but the symptoms and signs of renal failure resolve.

CHRONIC RENAL FAILURE—CRF is a loss of kidney function that occurs over a number of years. Azotemia is the ac-

cumulation of nitrogenous waste products in the blood caused by renal failure. Uremia refers to the symptoms and signs caused by CRF when renal function is less than about 10% of normal.

Etiology—Many diseases can destroy renal parenchymal tissue and result in CRF. These include chronic glomerulonephritis, hypertension, diabetes mellitus, polycystic kidney disease, analgesic nephropathy, nephrocalcinosis, reflex nephropathy, chronic pyelonephritis, obstructive uropathy, and interstitial nephritis. In certain patients, more than one disease may have caused the CRF. In some cases it is not possible to establish the cause.

Pathophysiology—CRF develops because the number of functioning nephrons decreases below that necessary to maintain homeostasis. Uremia and end stage renal disease (ESRD) occur when 90–95% of the nephrons are destroyed. As renal function deteriorates, hypertrophy occurs in the remaining nephrons and the amount of solute and water excreted per nephron may increase. Compensatory mechanisms eventually are overwhelmed by even the normal daily intake of water, sodium, potassium, acid, and nitrogen. Uremia, electrolyte disturbances, and fluid overload ensue.

The earliest renal impairment is the loss of ability to concentrate urine. This is due partially to the increased solute load per nephron. The patient then must increase water intake to prevent dehydration. The diurnal pattern of water excretion is reversed.

Most patients develop a tendency to retain salt and water early in the course of CRF. In a few forms of renal failure, salt wasting occurs because the kidneys are unable to conserve sodium even when sodium intake is restricted. The osmotic diuresis of the solute load causes an obligatory sodium loss. Hyponatremia and hypovolemia may occur and worsen renal failure by reducing the GFR. Salt-wasting eventually ceases and the kidneys are then unable to excrete dietary sodium. Sodium and water retention then results in edema and hypertension.

Serum potassium is normal during the early stages of renal failure. Renin-angiotensin-induced production of aldosterone stimulates potassium excretion and the osmotic diuresis further enhances potassium excretion. Eventually, the urine volume may fall below 500 ml/day, and serum potassium will begin to rise. Acidosis worsens hyperkalemia by causing the movement of potassium out of cells.

As renal function deteriorates, ability to form ammonia and therefore to excrete hydrogen is impaired. Ability to reabsorb filtered bicarbonate is also impaired. Acidosis ensues.

The percentage of phosphate excreted decreases as the GFR declines. The increased serum phosphate level and other factors described below cause a drop in the serum calcium level. Hypocalcemia stimulates the production of parathyroid hormone, which increases renal excretion of phosphate and resorption of calcium from bones. When the GFR reaches less than 20 ml/minute, the increased serum PTH level is no longer effective in increasing phosphate excretion.

Hypocalcemia is due to other factors besides the increased serum phosphate. Hypoalbuminemia reduces the quantity of carrier proteins for calcium. Absorption of calcium from the GI tract is impaired because of lack of 1,25-dihydroxyvitamin D, the active metabolite of vitamin D. The ionized fraction of serum calcium is decreased because ions such as sulfate, phosphate, and citrate bind the calcium.

Magnesium levels usually do not rise until the GFR is below 30 ml/minute. Uric acid levels rise, but not usually above 10 mg/dl, and gouty arthritis is uncommon.

Urea is poorly excreted in CRF, and the BUN rises. The magnitude of the rise correlates poorly with the symptoms of uremia except for the gastrointestinal symptoms. Increased quantities of urea are excreted into the intestinal lumen, presumably contributing to irritation and ulceration.

Other presumably toxic substances accumulate in uremia. These include indoles, phenols, amino acids, organic acids, and derivatives of guanidine. The accumulation of carotene-like pigments results in sallow skin color.

A normochromic normocytic anemia parallels the severity of the azotemia. Decreased RBC production occurs because erythropoietin deficiency and iron deficiency due to chronic GI blood loss. The anemia of chronic disease also is found in these patients. (See Hematology section.)

Several complications may occur. A bleeding tendency is caused by platelet dysfunction. The accumulation of guanidinosuccinic acid may be responsible for loss of platelet adhesiveness and aggregation. Osteomalacia occurs in part because vitamin D is not converted to the active metabolite, 1,25-dihydroxyvitamin D. Hypertension is exacerbated by fluid retention. A peripheral demyelinating neuropathy, mostly in the legs, results in decreased nerve conduction and impairment of motor and sensory function. Pericarditis may or may not cause chest pain and occasionally causes pericardial tamponade or constriction.

Renal-failure patients are predisposed to infections because of poor nutrition, pulmonary edema, lack of physical activity, vascular insufficiency, and indwelling tubes and catheters. Repeated transfusions increase the risk of viral hepatitis.

Symptoms and Signs—The onset of renal failure is insidious. The first symptoms may be polyuria or nocturia or both. Hypertension and anemia are common early signs, but lack specificity. As renal function deteriorates, the symptoms and signs relate to the organ systems involved.

Fluid accumulation produces the symptoms and signs of edema, congestive heart failure, and hypertension. Hyponatremia causes inability to concentrate, drowsiness, lethargy, psychotic disturbances, stupor, and coma. Hyperkalemia may cause cardiac arrhythmias. Acidosis contributes to nausea, fatigue, malaise, and dyspnea, and causes Kussmaul respiration. Hypocalcemia may result in tremor, muscle twitching, muscle cramps and convulsions. The increased PTH level leads to the erosive and cystic changes and bone pain of osteitis fibrosa cystica. Phosphate deposition in the skin contributes to severe itching; in the eyes, to conjunctivitis; in the blood vessels, to gangrene; and around the joints, to pain. Hypermagnesemia results in drowsiness, muscle weakness, and coma.

Ammonia formation from urea in the GI tract contributes to the unpleasant taste, anorexia, nausea, vomiting, and hiccups. Pericarditis may cause pain and be detected by hearing a friction rub. Pulmonary congestion from hypervolemia may cause dyspnea and hypoxemia. Urea in sweat precipitates on the skin and is known as "uremic frost."

The symptoms and signs of anemia are seen when the hematocrit falls below 15–20%. Patients with renal failure experience ecchymoses, epistaxis, and oozing of blood from mucous membranes due to coagulation abnormalities.

Neuropathy causes numbness, tingling, muscular weakness and, on occasion, paralysis.

The symptoms and signs of uremia progressively worsen. Renal failure is fatal unless the patient is treated by hemo- or peritoneal dialysis or receives a renal transplant.

Acid-Base and Fluid and Electrolyte Disturbances

Acid-base and fluid and electrolyte disturbances can be caused by a wide variety of diseases, including the kidney disorders previously discussed in this section. They also may be caused by gastrointestinal (eg, severe diarrhea), pulmonary (eg, chronic obstructive lung disease), or metabolic (eg, diabetes) disorders. The defects observed with these diseases have been described in earlier sections of this chapter.

Normal Physiology—A number of mechanisms act to maintain normal plasma pH (7.35 to 7.45), one of which is the chemical buffering by extra and intracellular buffer systems. These include hemoglobin, plasma proteins, and the carbonic acid-bicarbonate buffer system. Hydrogen ions ($H+$) migrate into or out of cells in exchange for potassium ($K+$) to maintain electrical neutrality. The respiratory system contributes through the exchange of carbon dioxide (an acid-former). Lastly, the kidneys help to maintain normal pH through the elimination or retention of $H+$ and bicarbonate (HCO_3^-). Each of these mechanisms acts to maintain a constant HCO_3^-:CO_2 ratio of approximately 20:1. As long as this ratio is maintained, the pH will be 7.4 (see Chapter 17).

The human body is composed largely of water. Fifty to 60% of total body weight is water. Body water is distributed between the intracellular space (intracellular fluid or ICF) and the extracellular space (extracellular fluid or ECF). Two-thirds of all body water is contained in the ICF and the remaining one-third in the ECF. The ECF is further divided into intravascular fluid (IVF) and interstitial fluid, which contain one-fourth and three-fourths of the ECF, respectively. Electrolytes are unequally divided between ICF and ECF. Potassium is the major ICF cation, and phosphate and organic ions are the ICF anions. Sodium is the major ECF cation, and chloride and bicarbonate are the ECF anions. Although water moves readily in and out of cells, electrolytes do not, often requiring active transport. Although electrolyte composition differs between the ICF and ECF, osmolality is equal.

Water homeostasis is regulated by the interrelationships between water intake, kidney function, and water loss through the lungs, skin, and GI tract. A decrease in ECF volume or an increase in osmotic pressure of plasma both stimulate water intake. The kidneys act to preserve water homeostasis through their relationship to antidiuretic hormone (ADH), which was discussed under **Endocrinology**. ADH release is under the control of both osmotic and volume factors. Increased osmotic pressure or decreased ECF volume stimulates increased ADH production and secretion. The glomerular filtration rate (GFR) is normally 125 ml/minute. The GFR is affected by renal blood flow, hydrostatic pres-

sure in Bowman's space, and plasma protein concentration. Essentially everything in the plasma, except protein, is filtered. The kidney tubules both reabsorb and secrete solutes via active transport and passive diffusion. Almost all water (90%) and electrolytes initially filtered are reabsorbed by active transport in the tubules and Henle's loop. Ammonia and urea are secreted into the filtrate.

Pathophysiology—Acid-base disorders may be divided into respiratory acidosis and alkalosis and metabolic acidosis and alkalosis. *Respiratory acidosis* is associated with disorders that cause an impairment of gas exchange and thus CO_2 retention. Arterial blood gases (ABGs) show a decreased pH, elevated pCO_2 (dissolved CO_2 gas) and elevated bicarbonate. *Respiratory alkalosis* is caused by conditions that result in hyperventilation with an abnormally large loss of CO_2. ABGs reflect an increased pH and decreased PCO_2 and HCO_3-. *Metabolic acidosis* occurs as a result of either the addition of acid or a loss of bicarbonate. Acids may be endogenous, as in the case of diabetic ketoacidosis, or exogenous, as in the case of methanol ingestion. Bicarbonate may be lost through diarrhea or through the kidneys as in renal tubular acidosis. ABGs show low pH, HCO_3-, and PCO_2. Calculation of the anion gap ($Na+ - Cl^- + HCO_3$-) is helpful in determining whether metabolic acidosis is due to addition of acid or loss of HCO3-. The normal anion gap is 10 to 12 mEq/L and is elevated when acidosis is due to addition of organic acid. *Metabolic alkalosis* usually is due to the loss of acid (H+) but may occur occasionally with excessive HCO_3- ingestion. Elevated pH and HCO_3- characterize it.

Once one of the above conditions occurs, the body compensates. For example, in cases of metabolic acidosis, the body compensates with increased respiratory activity, thus removing CO_2 and thereby blunting the fall in pH.

The causes of fluid and electrolyte imbalances are many. Such derangements may be interrelated, occurring together, or may occur independently. Fluid losses occur with gastrointestinal disorders such as vomiting and diarrhea. In such cases, electrolytes are lost with the water. In others, the losses of electrolytes and water are not proportional resulting in hypo- or hyperosmolality. In the various renal disorders, a number of fluid and electrolyte shifts are common. In the diuretic phase of acute tubular necrosis, large volumes of fluid are lost due to lack of reabsorption. In nephrotic syndrome, large shifts of water are often involved. This water is not lost necessarily from the body but may be lost from the vascular compartment, frequently in the form of edema. In addition to the fluid shifts, electrolyte disturbances ensue. Secondary to decreased renal blood flow and thus, decreased glomerular filtration rate, the renin-angiotensin system is activated causing further fluid retention. The specific renal diseases associated with fluid and electrolyte disturbances have been described in greater detail earlier in this section.

Symptoms and Signs—Signs of volume depletion include postural hypotension and tachycardia and decreased jugular venous pressure. Less reliable signs include decreased skin turgor, dry mucous membranes, and cloudy sensorium. Severe hypovolemia can result in shock. Fluid excess may be manifested by hypertension or peripheral or pulmonary edema. Of all electrolyte disturbances, only two of the more serious, those involving K+, will be discussed here. Others have been discussed in previous sections. Signs of hyperkalemia include muscle weakness and cardiac dysrhythmias. Severe hyperkalemia results in cardiac standstill. Hypokalemia also may be reflected as muscle weakness. Abdominal distress may occur from impaired intestinal smooth muscle mobility. Abdominal distention and depressed deep tendon reflexes may be evident. Cardiac rhythm disturbances also occur with hypokalemia.

The measurement of ABGs, plasma electrolytes, urine volume, and electrolytes are all helpful in assessing a patient with acid-base or fluid and electrolyte disorders, but the most helpful information frequently comes from careful physical examination (blood pressure, pulse rate, and jugular venous pressure).

NEUROLOGY

Epilepsy and Convulsive Disorders

Epilepsy is a chronic disorder of cerebral function. It may be defined as a paroxysmal disturbance of CNS function that is recurrent, stereotyped in character, and associated with excessive neuronal discharge that is synchronous and self-limited. The episodic manifestations of epilepsy are dependent on the portion(s) of the CNS involved.

Epidemiology—A total of 0.5–1% of the population suffers from epilepsy.

Epilepsy can begin at almost any age. However, the age of onset often is related to the etiology of the seizure disorder. One example is that of generalized absence seizures or petit mal, which typically present in early childhood.

Etiology—Epilepsy is a symptom complex that has many causes. In many cases, the precipitating factor(s) or cause of the seizure disorder is not apparent, and the condition is referred to as idiopathic. Severe hypoxia, genetic metabolic defects, developmental brain defects, and perinatal injuries can lead to seizures in newborns and infants. Certain metabolic disorders such as hypoglycemia, hypocalcemia, and vitamin B6 deficiency also can lead to seizures during infancy. Brain infections such as meningitis and encephalitis can trigger seizures during childhood. Seizures during childhood are less often caused by tumors, toxins, vascular disease, degenerative disease, or trauma.

In young adults, head trauma is a major cause of seizures. Likewise, ruling out the presence of a brain tumor is important for anyone over the age of 20. In patients over the age of 50, cerebrovascular disease is the most identified cause of seizures. In certain forms of epilepsy, genetic predisposition plays a role. Individuals with a first-degree relative with epilepsy are at a somewhat greater risk than the normal population of developing a seizure disorder. Despite our growing understanding of the disorder itself, an etiological diagnosis cannot be made with certainty in about two-thirds of epileptic patients.

In all age groups, a wide variety of drugs can provoke seizures.

Pathology—Various lesions in the brain, such as congenital lesions, gliotic scars, abnormal vascularization, and degenerative brain disease in the elderly have been associated with epilepsy in some patients and not in others. Even when the clinical information suggests that a seizure is of focal origin, it is not always possible to identify the epileptogenic lesion.

Pathophysiology—The convulsion results from sudden hypersynchronization of electrical discharge in neuronal networks in an apparently normal or a diseased cerebral cortex. The mechanisms and reasons for the discharge are not well understood. One hypothesis is that a group of diencephalic neurons normally exerts a constant inhibitory influence on cortical neurons, thereby preventing excessive discharge. In epilepsy, the neurons are deafferented, supersensitive, and susceptible to activation or depolarization by a variety of stimuli. Seizures may result from a reduction of inhibitory neurotransmission mediated by the neurotransmitter gamma-aminobutyric acid (GABA) or by enhancement of the excitatory neurotransmitter system mediated by glutamate and aspartate.

During a seizure, consciousness may be unaffected, lost completely, or altered but not completely lost. Patients may experience only minor interruptions in their motor activity or they may experience intense muscular activation that leads to motor behavior characteristic of generalized tonic-clonic seizures.

Symptoms and Signs—The frequency of seizures within individual patients can vary from as few as one per year to dozens per day depending on the particular seizure type. Thus, the need for accurate diagnosis of the seizure type is of more than just theoretical interest. Since there are so many different seizure types, each of which may require a different therapeutic approach, an accurate diagnosis permits the clinician to select the most appropriate anticonvulsant drug while avoiding the use of contraindicated drugs. The International Classification of Epileptic Seizures classifies seizure types as either partial or generalized.

Partial seizures generally are categorized as simple, complex, or secondarily generalized and would include those traditionally called focal motor and temporal lobe seizures. Partial seizures all begin from a discrete brain region and may or may not be preceded by an aura. An aura consists of sensations or experiences often recognized by the patient as a warning of an impending seizure. Partial seizures may or may not involve loss of consciousness.

The symptoms of simple partial seizures result from abnormal discharges originating in specific areas of the cortex, and often remain unilateral regardless of whether the seizure is motor, somatosensory, psychic, autonomic, or a combination. The aura of a simple partial seizure may include somatosensory symptoms or hallucinations (eg, tingling, light flashes, or buzzing); autonomic symptoms including epigastric sensation, pallor, sweating, flushing, piloerection, and pupillary dilation; or psychic symptoms.

One form of a simple partial seizure is that traditionally known as *Jacksonian*. It usually begins with twitching of the fingers of one hand, the face, or one foot. The movement then spreads (marches) to other muscles along the same side of the body. If the movements generalize to include both sides of the body and the patient loses consciousness, the seizure is said to have become secondarily generalized. One type of a partial seizure with complex symptomatology is traditionally known as a psychomotor seizure.

In a complex partial seizure, consciousness is lost. Complex partial seizures are associated often with a lesion in the temporal lobe. The patient acts as though he/she were conscious, although he/she is amnesic. The patient may continue an activity or perform tasks but may not be able to respond to questions or commands. The seizure often is preceded

by an aura. Motor activity due to the seizure may include chewing, lip-smacking, and tonic spasms of the extremities.

Generalized seizures, on the other hand, involve both hemispheres from the beginning. Consciousness is lost at the outset, and patients experiencing generalized seizures usually do not experience aura or display focal motor manifestations. The two most widely recognized generalized seizure disorders include generalized tonic-clonic (formerly grand mal) and generalized absence (formerly petit mal). Generalized tonic-clonic seizures are characterized by a sudden loss of consciousness, a cry, falling, tonic then clonic movements of the muscles, and incontinence of sphincters. After the motor seizure has ceased, the patient may be unconscious for many minutes. On awakening, the patient may complain of a headache. Generalized absence seizures almost always begin between 4 and 12 years of age. They are characterized by a brief loss of consciousness lasting for a few seconds. The child typically displays a blank facial expression and may or may not display a characteristic blinking of the eyelids. Absence seizures are almost always associated with a typical EEG abnormality of spike and slow wave discharges of approximately 3 Hz. Other generalized seizures include myoclonic seizures, clonic seizures, tonic seizures, and atonic seizures.

The International Classification of Epilepsies and Epileptic Syndromes takes into consideration the fact that some patients with epilepsy display more than one seizure type. After all, seizures are only a symptom of the underlying disorder. The prognosis is often a product of the epileptic syndrome whose diagnosis depends on numerous factors including family history, age of onset, rate of progression, presence or absence of neurological impairment and interictal EEG abnormalities, and a patient's response to pharmacological treatment. In this respect, the epileptic syndrome generally is classified according to whether an individual patient's seizures are localization-related (focal, local, partial) or generalized and whether they are idiopathic or symptomatic. They may be classified further as to anatomical localization (eg, frontal lobe, Rolandic, occipital, or temporal epilepsy). To date, more than 50 epileptic syndromes have been proposed.

The diagnosis of epilepsy is based on the clinical history and the EEG. The first steps to an accurate diagnosis usually involve obtaining an accurate and complete history from the patient as well as from a witness. A detailed physical exam is followed by an even more in-depth neurological exam. The EEG can provide precise information that may be useful for classifying the seizure type. It is characteristically abnormal during a seizure but may be normal between seizures. Specialized diagnostic procedures may include computerized tomography and magnetic resonance imaging. These two noninvasive techniques can be useful in identifying a particular brain lesion that may have led to the development of a patient's seizure disorder. Intensive monitoring employing closed circuit television and EEG recording is an expensive procedure that should be considered when a patient's seizures are not responsive to drug therapy. This latter procedure also can be helpful in determining whether difficult-to-diagnose seizures are nonepileptic in nature.

Parkinsonism

This disease, also called paralysis agitans, is a disorder of the extrapyramidal system that originally was described in 1817. James Parkinson described a syndrome that consisted of a resting tremor, rigidity, postural abnormalities, and bradykinesia, but spared the senses and intellect.

Normal Physiology—The basal ganglia normally control postural tone and provide the background adjustments for intentional movements. The dopaminergic pathway from the caudate nucleus to the thalamus inhibits the inhibition of voluntary movement. The cholinergic pathway opposes this pathway, which is excitatory for the inhibition of voluntary movement. A dopaminergic pathway inhibits the cholinergic pathway in the caudate nucleus from the substantia nigra.

Epidemiology—Parkinsonism usually occurs in middle or late life, though it rarely is seen in young people. The prevalence of this disease is estimated to be between 59 and 353 cases per 100,000 in various populations worldwide, resulting in 300,000 to 400,000 cases in this country.

Etiology—Although the actual cause of Parkinson's disease remains undetermined, there is accumulating evidence that multiple genetic and environmental factors interact to cause damage to extrapyramidal dopamine neurons. It is believed that 70–80% of these neurons must be lost before symptoms of Parkinson's disease become evident. Many of the persons who survived the pandemic of von Economo encephalitis in 1918 and 1922 developed Parkinsonism 20 to 30 years later. Psychoactive drugs such as the phenothiazines and butyrophenones can cause a syndrome similar to Parkinsonism. Infections, tumors, and certain chemicals and drugs may cause an identical but reversible disorder. The term Parkinson's disease is reserved for paralysis agitans of unknown cause. Autosomal recessive and sporadic juvenile cases are often caused by mutations in the parkin gene located at 6q25.2–27.

Pathology—Melanin is lost from nerve cells in the brainstem, particularly in the substantia nigra, and accompanied by extensive loss of dopaminergic nerve cells and reactive gliosis. Intracytoplasmic inclusion bodies, Lewy bodies, also can be found in surviving neurons in the affected areas.

Pathophysiology—Loss of inhibition and the unbalance of opposing pathways in the thalamus and caudate nucleus result in the movement difficulties of Parkinsonism. The origin of the tremor is less clear. Decreased dopamine is found in the substantia nigra, caudate nucleus, and putamen.

Symptoms and Signs—The clinical features are characteristic. There is often a prodromal phase consisting of nonspecific symptoms, such as fatigue, musculoskeletal pain, declining performance, and depression. Within 1 to 2 years, more definitive symptoms appear. The typical tremor occurs at rest and lessens with voluntary movement. The tremor may involve the hands, legs, lips, tongue, and eyelids when the eyes are closed. In the early stages of the disease, the tremor is unilateral but becomes bilateral later in the course. The tremor occurs at a frequency of 4 to 8 cycles/second. The hand tremor is described as "pill rolling."

In the early stage of the disease, there is bradykinesia as all movement is slowed. Later, the patient has particular difficulty initiating movement. Finally, there is absence of movement or akinesia. The spontaneous movements of posture change, such as arm swinging while walking, disappear. The face becomes expressionless and is known as mask-like facies. The voice becomes monotonous. The posture is stooped. Because the patient cannot make reflex adjustments to the posture changes of walking, "he walks with quick shuffling steps at an accelerating pace, as if attempting to catch up with his center of gravity." Passive movement of the extremities elicits "lead pipe rigidity" because both flexors and extensors are contracted or "cog wheel" motion from the superimposition of tremor on rigidity.

Anxiety and tension aggravate the symptoms. The patient also may have seborrhea, excessive sweating, and salivation.

Eventually the patient is incapacitated by the rigidity, and the tremor disappears. The clinical course is one of gradual deterioration.

Stroke Syndromes

A stroke is a process involving one or more blood vessels in the brain, which results in the sudden and dramatic development of a focal neurological deficit. The deficit reflects the location and size of brain injury. Three separate entities are recognized: transient ischemic attack (TIA), stroke in evolution, and completed stroke. While TIAs are transient, evolving and completed strokes are not.

Etiology—The vast majority of strokes are caused by atherosclerotic thrombosis of the cerebral arteries. Embolism from the heart or ulcerated atherosclerotic plaques in the carotid arteries also causes them. Cerebral hemorrhages are due most often to hypertension but also may be due to the rupture of an aneurysm. Less frequent causes include trauma, excessive anticoagulation, hypercoagulable disorders, and inflammatory diseases of cerebral blood vessels.

Normal Physiology—The effects of the blood vessel occlusion relate to the location and availability of collateral or anastomotic blood flow. The circle of Willis provides collateral circulation and protects the brain from ischemia that would otherwise result from occlusion of a carotid or vertebral artery. Retrograde flow from the external carotid may prevent damage when the internal carotid is occluded. Collaterals for the vertebral artery exist. Other anastomoses may prevent or lessen damage if the lesion is distal to the circle of Willis.

Pathophysiology—Infarction results from occlusion of arteries of the brain as elsewhere in the body (see the discussion of atherosclerosis). Thrombotic stroke results when a thrombus develops on an atherosclerotic plaque: the lumen of the vessel is narrowed or may be occluded completely, and collaterals are insufficient to preserve function. Extension of the thrombus may block collateral blood flow. Dural sinus (venous) thrombosis may cause hemorrhagic infarction.

Cerebral embolism most commonly originates from a thrombus in the heart, particularly during atrial fibrillation. Other sources of embolic strokes are mural thrombi that occur after myocardial infarction and pieces of intra-arterial thrombi. The emboli usually lodge at bifurcations.

Intracranial hemorrhage is the third most frequent cause of stroke. Intracranial hemorrhage is due most commonly to hypertension, rupture of saccular aneurysm, and bleeding disorders. Cerebral hemor-

rhages due to hypertension involve a penetrating artery and occur within the brain tissue. Adjacent tissue is compressed and displaced by the mass of blood. Saccular aneurysms or berry aneurysms are thin-walled blisters protruding from the arteries of the circle of Willis or major branches of the circle at bifurcations. Developmental defects in the media of the arteries cause the aneurysms, which are composed of intima and adventitia. The defect in the wall structure is congenital, but enlargement and eventual rupture occur during later life, reaching a peak at 35 to 65 years. Rupture of the aneurysm results in bleeding into the subarachnoid space and occasionally into the brain as well.

Symptoms and Signs—The location of the lesion determines the nature of the deficit. Lesions in the carotid system result in unilateral signs of hemiplegia, hemihypoesthesia, hemianopia, aphasia, and agnosia. Lesions in the basilar system result in bilateral signs, motor and sensory deficit, brainstem deficit, and variable cranial nerve abnormalities. Cerebellar infarction results in severe dizziness, nausea, vomiting, ataxia, and nystagmus.

In most cases of thrombotic stroke, a TIA has occurred previously. A TIA due to temporary or partial occlusion of all or part of the carotid or middle cerebral artery system may consist of hemiplegia, hemiparesthesia, monocular blindness, or other focal signs, depending on the area of brain affected. A TIA due to temporary or partial occlusion of the vertebral-basilar system consists of dizziness, diplopia, numbness, impaired vision, and dysarthria. A TIA usually lasts for about 10 minutes but may last from a few seconds up to 24 hours. Between the TIAs, the patient may have no neurological deficit. A bruit may be heard over the carotid arteries if they are severely atherosclerotic.

A thrombotic stroke begins suddenly but may progress over several days. Parts of the body may become involved in a stepwise fashion. A completed stroke is defined as 18 to 24 hours without progression for the carotid system and 72 hours without progression for the vertebral-basilar system.

Prognosis in a thrombotic stroke is difficult to predict. Comatose patients have a poor prognosis. Improvement generally occurs as functions are taken over by other parts of the brain or when edema surrounding an infarct subsides. If improvement has not begun by the second week, prognosis is poor. Any deficit that remains at the end of 6 months is likely to be permanent.

Embolic strokes develop the most rapidly and are fully developed within minutes. No warning symptoms precede an embolic stroke. Focal deficits such as motor aphasia, receptive aphasia, or a sensorimotor paralysis may occur. The ultimate prognosis depends upon the correction of the underlying disease.

Cerebral hemorrhage due to hypertension occurs without warning and evolves over hours. It occurs more commonly and at a younger age in blacks. The symptoms and signs depend on the site and size of the hemorrhage. Hemorrhage is most common in the putamen, where it causes hemiplegia, hemisensory loss, homonymous visual loss, and aphasia when the lesion is on the dominant side. Severe headache and vomiting occur at the onset. A total of 85% of patients with cerebral hemorrhages due to hypertension do not survive the first 8 hours. A CT scan reliably detects intracerebral and intracerebellar hemorrhages of 1 cm or more if the study is performed within 2 weeks of the hemorrhage.

Rupture of a saccular aneurysm may present with sudden unconsciousness with or without preceding excruciating headache. There are no lateralizing neurological signs when the blood is confined to the subarachnoid space. The hemorrhage tends to recur if surgical correction is not carried out or is unsuccessful. Prognosis is poor if the patient is comatose; however, if the patient awakes, recovery is likely.

Headache

The three major types of primary headaches are migraine, cluster, and tension-type headaches. Migraine is divided into migraine with and without aura, (formerly called classic and common migraine respectively).

Epidemiology—Migraine affects about 12% of the population. It is four times as common in women, and frequently runs on families. Tension-type headache is several times more common than migraine.

Pathogenesis and Pathophysiology—The pathogenesis of migraine has been exhaustively investigated, but remains incompletely understood. An abnormality of the trigeminal nerve vasculature provoked by release of nitric oxide appears central, but other mediators, such as serotonin likely also play a role.

Tension-type headaches are not a result of scalp or neck muscle contraction. Their pathogenesis is debated but shares some features with migraine.

Symptoms and Signs—Migraine with aura occurs in three stages, prodrome, aura, and headache. The first stage or prodrome lasts from hours to days and may involve changes in mood or appetite and fluid retention. The prodrome may be unapparent or not recognized. The aura is most commonly visual (blurred or cloudy vision, scotomas, and/or flashes of light), but vertigo, chills, tremors, unilateral numbness, aphasia, photophobia, or pallor also may occur. As the aura subsides, the patient experiences a severe, throbbing headache, which initially is unilateral in most cases. Nausea, vomiting, diarrhea, chills, tremors, and perspiration also may occur at this time. During recovery the pain decreases markedly, but the head is tender and exhaustion is present. The migraine without aura lacks the aura phase, but the actual headache may last longer (more than 2 hours) than in migraine with aura.

Cluster headaches are usually unilateral, nonthrobbing, and are more common in males. The patient experiences excruciating pain lasting 20 to 90 minutes. Autonomic features such as nasal stuffiness, rhinorrhea, tearing, and pupillary changes frequently accompany the pain. Bouts may follow one another several times a day for 4 to 8 weeks, not to recur again for 6 to 12 months,

Tension-type headaches may cause intermittent, recurrent, or constant pain. Patients may describe scalp soreness with pain on combing their hair, band-like pain or tightness, and pressure.

Neuromuscular Disease

GUILLAIN-BARRE SYNDROME

This acute inflammatory demyelinating polyneuropathy results in flaccid paralysis with spontaneous recovery.

Epidemiology—Annual incidence is 0.6 to 2.4 cases/100,000/year. It is now the most common cause of flaccid paralysis.

Etiology—The cause of most cases of Guillain-Barre syndrome is unknown. Most cases follow within 3 months of an acute respiratory or gastrointestinal illness, most commonly campylobacter. Cases have been associated with many other infections or medical illnesses.

Pathophysiology—Pathological changes observed in patients who die of Guillain-Barre syndrome include perivascular lymphocytic infiltrates usually associated with demyelination of the affected nerves. Infiltrates also may occur in the liver, spleen, lymph nodes, and heart. Although the pathogenesis is unclear, the syndrome may involve a cell-mediated immunological reaction directed at peripheral nerves.

Symptoms and Signs—The principal symptom is muscle weakness of both proximal and distal limbs. The weakness may advance to muscles of the trunk. While loss of sensation is unusual, paresthesias often occur. Affected patients are afebrile. In severe cases, the respiratory system may be affected, requiring ventilator support. Death is rare, and complete recovery occurs in the majority of cases. Examination of the cerebrospinal fluid shows few cells but a distinct elevation in CSF protein. Nerve conduction studies show slowed motor nerve conduction with temporal dispersion and prolonged distal latencies.

MYASTHENIA GRAVIS

This is a disease characterized by muscle fatigability and weakness most prominently affecting the muscles of the eye and cranium.

Incidence and Epidemiology—The incidence is 1 in 20,000 in the general population. All age groups are affected with females in the 20- to 40-year age group predominating.

Etiology and Pathophysiology—While the underlying cause remains a mystery, the physiological defect has been clarified. The disease is due to a reduction in number and effectiveness of acetylcholine receptors at the motor and plate. This reduction is secondary to an autoimmune mechanism that destroys the receptors. In experimental models, massive phagocytic infiltration of motor end plates with large areas of postsynaptic membrane destruction and associated decrease in acetylcholine receptors is observed. This process results in the denervation of muscle fibers. There are other forms of myasthenia that are not associated with disturbed immunity, including inherited deficiencies in biosynthesis of acetylcholine or its receptors.

Symptoms and Signs—The typical clinical presentation includes drooping eyelids, aphasia, and the inability to perform usually simple muscular functions. Early in the disease, only a few muscles are affected. Neuromuscular fatigue is a cardinal sign: patients are unable to sustain or repeat muscular movements.

Electromyography is a useful diagnostic technique and shows a rapid decline in the amplitude of muscle action potentials with repetitive muscle contraction. Other tests used in diagnosis include the use of

anticholinesterase agents and the detection of antibodies to acetylcholine receptors, which can be demonstrated in 90% of patients with myasthenia gravis.

MULTIPLE SCLEROSIS

A number of neurological disorders are characterized by the degeneration of the myelin sheath of nerve fibers. Of these, only multiple sclerosis (MS) will be discussed. Other dieases falling into this classification include acute disseminated encephalomyelitis (postvaccinial and postinfectious encephalomyelitis) and acute hemorrhagic leukoencephalitis.

Normal Physiology—Many of the nerve fibers of the body are covered with a layer of lipid material called myelin. This myelin sheath is interrupted at intervals by spaces termed nodes of Ranvier. Myelinated nerves are found in great number in cranial and spinal processes and in the white matter of the brain and spinal cord. The myelin sheath facilitates rapid nerve impulse conduction.

Etiology and Epidemiology—The etiology is unclear although several epidemiological factors may offer some clues. This disease is rare between the equator and latitudes 30° to 35° north and south. It occurs more frequently with increasing latitude. MS is more common in some families, suggesting simultaneous exposure to some etiological agent or perhaps a hereditary factor. These factors suggest, to some, an infectious etiology with a resultant autoimmune response.

Pathophysiology—The pathologic lesions vary in size and appearance but always include or reflect demyelination. The lesions ("plaques") occur throughout the white matter of the CNS. They are located most commonly in subpial and periventricular white matter of the cerebrum, optic nerves, cerebellum, brainstem, and spinal cord. The associated pathophysiological change is a decrease in speed of nerve impulse conduction. Symptoms worsen with age, reflecting the ongoing nature of the disease.

Symptoms and Signs—While most patients present with evidence of spinal cord or brainstem involvement, about 40% present with only optic neuritis. The former presentation may include paresthesias, numbness, or weakness in an asymmetrical distribution. Diplopia, nystagmus, and cerebellar ataxia also may occur. The latter presentation may include partial or complete blindness in one or both eyes, scotomas, or pain with eye movement. This disease progresses with time with interspersed exacerbations and eventually may result in quadraplegia and coma. The usual patient survives 20 years or more from the time of the initial diagnosis. Magnetic resonance imaging (MRI) scans are the most sensitive means of detecting lesions. Cerebral spinal fluid may contain oligoclonal bands, myelin basic protein, or elevated IgG. Visual, somatosensory, or brainstem auditory evoked potentials may be abnormal and assist in the diagnosis.

Dementia

This is a generic term referring to a syndrome of declining cognitive function. The clinical course of the disorder is extremely variable, and the causes are probably multiple. About 70% of progressive dementias are believed to be due to Alzheimer's disease (AD).

Pathophysiology—Many types of dementia involve structural disease of the cerebrum and diencephalon. Degeneration and loss of nerve cells with secondary changes in the cerebral white matter often are observed. These changes may occur in one or many parts of the brain. AD is characterized by neurofibrillary tangles and senile plaques, found prominently in the hippocampus and association cortex. While the underlying etiology is often undetectable, dementia with its various lesions may be due to identifiable disorders such as chronic hydrocephalus, syphilis, and certain virus infections.

Symptoms and Signs—The initial presentation is quite variable. Symptoms include irritability, lack of interest, distractibility, unclear thinking, loss of memory, and wide mood swings. As the disorder progresses, incontinence, aphasia, and speech disorders often develop. Eventually, the patient becomes unable to care for himself and apparently has no interest in doing so. The course is variable with progression occurring over months or years. It should be stressed that the disease may be due to a wide variety of disorders, many of which are treatable. Therefore, a detailed diagnostic effort is warranted.

RHEUMATOLOGY

Normal Physiology—Joints allow movement of one bone upon another. The ends of the bones are covered with hyaline cartilage, and diarthrodial joints are covered by collagenous tissue called the joint capsule. The synovial membrane lines the joint space side of the joint capsule. The synovial membrane is a relatively acellular, highly vascular, delicate membrane that secretes the synovial fluid. Cartilage, which is avascular, derives its nutrition from the synovial fluid. Various inflammatory diseases, trauma, and degeneration may involve the joint.

Rheumatoid Arthritis

Rheumatoid arthritis (RA) is a systemic autoimmune disorder of unknown etiology. It is characterized by chronic, symmetric, and erosive destruction of the peripheral joints. The severity of the joint disease may fluctuate over time, but generally, joint destruction and deformity are the end results of this disease. There are also common manifestations of this disease affecting other body systems. For instance, subcutaneous nodules, pulmonary nodules and fibrosis, vasculitis, pericarditis, and episcleritis of the eye, are just some of the examples of extra-articular involvement.

Epidemiology—Approximately 3 million people in the US have RA. The onset is most common in the 3rd and 4th decades but may affect all age groups, including children. Women develop the disease more commonly than men do by a ratio of 2.5:1. The prevalence of disease increases with age for both males and females.

Etiology—The etiology is unknown. Histocompatibility typing has proven that a predisposition for the disease is inherited. Unknown environmental factors may play a role in the development of RA. Viruses and bacteria are suspected as possible causes, although to date there is no convincing evidence to support their etiologic role.

Pathophysiology—The disease is characterized by inflammation of the synovium. Infiltration by mononuclear leukocytes occurs along with edema, vascular congestion, and fibrin deposition. As a result of chronic inflammation, the synovium thickens, and forms large villi, which protrude into the joint space. This is referred to as a pannus. The pannus erodes the underlying cartilage and bone. The inflammatory process and destruction of normal joint anatomy results in weakening of tendons, ligaments, and other supporting structures. This leads to instability and partial dislocation (subluxation) of the joint.

Rheumatoid nodules, characteristic of RA, are found most commonly in subcutaneous tissue over pressure points such as the extensor surface of the forearms. However, they also may be found in the lung, heart, or vocal cords. Microscopically, the nodules contain a central area of necrosis surrounded by palisading epitheloid cells and chronic inflammatory cells. Severe RA also may be complicated by vasculitis involving multiple organs.

Antibodies against immunoglobulin G (IgG) are found in the serum and synovial fluid of most patients with RA. The antibodies are of the IgM, IgG, and IgA classes of immunoglobulins and are called rheumatoid factors. Chronic antigenic stimulation is thought to induce production of these antibodies. The exact role of rheumatoid factor in the development of RA has not been demonstrated. However, immunologic mechanisms do appear to play a role in the pathogenesis of RA. Immune complexes of immunoglobulins, rheumatoid factor, and complement generate vasoactive and chemotactic substances in the joint. Lysosomal enzymes, which cause tissue injury, are released after phagocytic cells ingest the immune complexes. It is the release of these vasoactive substances and enzymes that are primarily responsible for the joint erosion and destruction that characterizes this disease.

Symptoms and Signs—The onset is usually insidious. Fatigue, weakness, joint stiffness, arthralgias, and myalgias may precede signs of joint inflammation. The joints gradually become tender, swollen, hot, and painful. Joint stiffness, particularly after a prolonged period of rest ("gelling"), is a major complaint of patients with RA. Morning stiffness is a particular and almost universal complaint of patients with RA. In contrast to the rather brief (5–10 minutes) of gelling seen in patients with osteoarthritis, the morning stiffness of RA is prolonged, sometimes lasting in excess of 1 hour.

Nearly all patients with RA will have synovitis of the wrist, metacarpophalangeal joints (MCP), and proximal interphalangeal joints (PIP) of the hands. Typically the distal interphalangeal joints (DIP) are spared. The cervical spine is frequently involved but interestingly, disease of the thoracic and lumbar spine is exceptionally rare. Other commonly affected joints are the shoulders, elbows, hips, knees, ankles, and metatarsophalangeal joints (MTP) of the feet.

The hypertrophied synovium of involved joints may be palpated. Muscle weakness and atrophy often parallel the severity of the joint disease. Range of motion, especially extension, becomes limited and can lead to flexion contractures. Swan-neck, boutonniere, and cock-up toes are terms used to describe the deformities of the hands and feet. Ulnar deviation of the fingers can occur.

Duration of morning stiffness, which usually is measured in hours, may be used to monitor disease activity. Other indicators include grip strength, time required to walk a certain distance, number and clinical assessment of joints involved, and radiographs demonstrating erosion of bone, loss of joint space, and soft-tissue swelling.

RA is a systemic disease involving multiple organ systems besides the joints. Rheumatoid nodules are found in 20% of RA patients. Less than 5% of the patients have vasculitis. However, the vasculitis may be severe and can result in peripheral neuropathy, nail-fold thrombi, digital gangrene, and leg ulcers. The most common ocular manifestation is keratoconjunctivitis sicca (Sjögren's syndrome); episcleritis also may occur. In the lungs, interstitial fibrosis, rheumatoid nodules, and pleural effusions are seen. Inflammation of the pericardium may cause pericarditis. Rarely this may result in cardiac tamponade. Rheumatoid nodules on the heart valves may lead to murmurs and nodules in the heart muscle that can cause electrical conduction disturbances.

Patients with severe arthritis may develop Felty's syndrome. Felty's syndrome was originally described as RA, splenomegaly, leukopenia, and leg ulcers. However, subsequent observations have shown that there is an additional association with lymphadenopathy and thrombocytopenia.

Mild to moderate anemia that is normochromic or hypochromic is found in patients with RA. The severity of the anemia parallels the activity of the disease. The defect is thought to be in iron utilization in hemoglobin synthesis (see anemia of chronic disease).

Other abnormal laboratory tests include a high erythrocyte sedimentation rate, which may be used to monitor disease activity. The latex aggregation test for IgM rheumatoid factor is positive in 70–80% of patients. Unfortunately, other diseases of chronic inflammation also are associated with a positive rheumatoid factor test, therefore, it is not specific to RA despite its name. Analysis of the synovial fluid, while not diagnostic, typically shows neutrophils (10,000–50,000/mm^3) and elevated protein levels.

Diagnosis—The highly variable clinical course of RA makes prognosis difficult in individual patients. Spontaneous remissions and exacerbations are characteristic. Remissions occur most frequently in the early stages of the disease. Some patients may experience a complete remission with little or no joint deformity. Others have a chronically progressive course over many years with development of varying degrees of joint damage. A smaller group, 10–15%, has a relentless destructive course that results in severe deformities and crippling. The unpredictable course of RA also makes evaluation of therapy particularly difficult and contributes to the quackery seen in this field.

The diagnosis of RA is based on the clinical picture of symmetrical inflammatory arthritis usually involving small joints, characteristic radiograph changes, and a positive rheumatoid factor test. Other causes of inflammatory arthritis are Reiter's syndrome, psoriatic arthritis, and systemic lupus erythematosus. Arthritis associated with inflammatory bowel disease must be excluded. The arthritis associated with Lyme disease or hepatitis B may mimic RA. Degenerative joint disease may occur simultaneously.

Degenerative Joint Disease

Loss of joint cartilage and hypertrophy of bone characterize degenerative joint disease (DJD), also known as osteoarthritis (OA).

Epidemiology—Approximately 40 million Americans have radiographic evidence of DJD, but many have no symptoms attributable to the disease. The prevalence of DJD increases with age, 85% of people 75 years or older have characteristic radiographic changes. DJD is a major cause of disability. Severe osteoarthritis of the knee is more likely to result in disability than significant involvement of any other joint.

There are racial and gender differences in both the prevalence and pattern of joint involvement for DJD. For example, Caucasians have a higher rate of hip osteoarthritis than do Blacks, Native Americans, or Asian races. Women are twice as likely as men to have OA of the knees, and Black women twice as likely as Caucasian women.

Etiology—Evidence indicates that heavy use of a joint, so-called "wear and tear" may play a role in initiating the degeneration of cartilage. In other patients, degenerative changes occur when infection, acute trauma, excessive use, or congenital deformities have damaged the cartilage. The precise mechanisms of cartilage loss in DJD are unknown. Obesity has also been linked to increased prevalence of OA, especially of the knee. Likewise, genetic factors seem to have additional roles in the development of DJD. For instance, in a woman with DJD of her distal interphalangeal joints (Heberden's nodes), her mother is twice as likely and her sister three times as likely to have the same findings than the mother or sister of an unaffected woman. The mechanism of transmission appears to be autosomal dominant in women and recessive in men.

Pathophysiology—Degenerative joint disease essentially develops in two settings: when there is normal cartilage and bone but abnormal stress or excessive loads placed on the joint which cause the tissues to fail; and when there is a normal applied stress but the underlying joint tissues are defective. DJD may be classified as either primary or secondary. No predisposing cause can be identified in primary DJD. Causes of secondary DJD include infection, trauma, fractures, unusual use, and damage by inflammation as in RA, and congenital abnormalities. In addition, acromegaly, alkaptonuria, hemochromatosis, and chrondrocalcinosis are predisposing factors for secondary DJD.

In either case, histologically degenerative changes are seen in cartilage as progressive loss of metachromasia, which is evidence of proteoglycan loss. Chondrocytes increase in number and form clusters. The surface of the cartilage loosens, flakes off, and fissures form as deeper layers become involved. The cartilage may be lost completely. The bone at joint margins responds by osteophyte formation and hypertrophy. The subchondral bone, which has lost the covering cartilage, becomes dense, smooth, and glistening (eburnation). Cystic areas may develop below the joint surface. Inflammation of the synovium and joint capsule is usually mild.

Collagen fibers and proteoglycans give normal cartilage the properties of compressibility and elasticity. The proteoglycan molecules bind large numbers of water molecules that are released when the cartilage is compressed and are regained when the force is removed. The proteoglycan content of DJD cartilage is diminished and the molecular species is altered.

In contrast to normal adult cartilage, the chrondrocytes proliferate. The chrondrocytes are continuously rebuilding the cartilage matrix in DJD. The amount of hydrolases is increased. As the disease progresses, the destruction exceeds the rate of repair, resulting in a net loss of cartilage. Cartilage laid down during the rebuilding process is of the type normally found in tendons and skin but not in bone. Simultaneously, the subchondral bone sclerosis and marginal bone overgrowths (spurs) develop.

Symptoms and Signs—Pain in the joints particularly with motion or weight bearing is characteristic of DJD. Joint stiffness occurs after rest and quickly subsides after resuming movement. The duration of morning stiffness is measured in minutes rather than hours as in RA.

Examination of the joints reveals decreased range of motion, local tenderness, bony enlargement, but usually no heat or erythema. DJD commonly involves the distal interphalangeal (DIP) joints, in contrast to RA. Bony enlargement of the DIP joints is called Heberden's nodes. Enlargement of the proximal interphalangeal (PIP) joints is known as Bouchard's nodes. DJD involvement of the spine may cause compression of spinal nerve roots by the bony spurs, which can lead to a variety of complaints. DJD of the hips and knees may be the most disabling form of the disease.

Diagnosis—There is no laboratory abnormality characteristic of DJD. The diagnosis is based on symptoms and signs and the radiographic changes of joint space narrowing and bony spur formation.

Crystal-Induced Arthritis: Gout and Pseudogout

Several distinct diseases are characterized by crystal deposition in and about joint spaces. This deposition can lead to acute inflammation of the joint. Gout is a disorder of sodium urate deposition whereas pseudogout is characterized by deposition of calcium pyrophosphate dihydrate crystals. Recently, a form of arthritis has been attributed to hydroxyapatite deposition.

Epidemiology—Contrary to folklore, gout is not related to socioeconomic class. Few individuals with gout consume excessive quantities of purine-containing foods. Primary gout is a disease primarily of the adult male with a peak incidence in the 5th decade. Only 10–15% of cases occur in females, and these are usually in the postmenopausal group. Hyperuricemia is found in 5% of all asymptomatic persons at least one time during adulthood. However, fewer than one in five will develop clinically evident crystal deposition.

Diabetes mellitus, obesity, hypertension, coronary and cerebral atherosclerosis, and hypertriglyceridemia all occur more frequently among gouty patients for unknown reasons.

Etiology—The etiology of gout is either the overproduction or the underexcretion of uric acid. Overproduction may be primary and due to enzyme deficiencies in the metabolic pathway for purines; or may be secondary due to increased purine turnover as in hemolytic or myeloproliferative diseases. Occasionally, increased dietary consumption may cause increased levels as can ethanol abuse. Uric acid underexcretion may be caused by diminished renal function, interaction with various medications, or may be idiopathic.

Calcium pyrophosphate dihydrate crystal deposition disease (CPPD) is due to hereditary causes, trauma, or may be associated with certain metabolic diseases such as hemochromatosis, hypothyroidism, hyperparathyroidism, and amyloidosis,

Pathophysiology—The rates of production and elimination of uric acid determine the amount of uric acid in the body. Exogenous (dietary) and endogenous purines are oxidized to uric acid and eliminated. Of the uric acid eliminated, the kidney excretes two-thirds and the gastrointestinal tract excretes the remainder. The two most important processes in the development of hyperuricemia are abnormalities of endogenous purine production and of uric acid excretion by the kidneys. The majority of patients with gout have a defect of uric acid clearance through the kidneys. Specific enzyme abnormalities that have been identified include decreased hypoxanthine-guanine phosphoribosyltransferase and increased PP-ribose-P synthetase, which result in the overproduction of uric acid.

Uric acid is filtered by glomeruli, but 98% of the filtered amount is reabsorbed by the tubules. The majority of the uric acid excreted (80–85%) is secreted actively into the urine by the renal tubules. The exact reason for undersecretion of uric acid by the tubules is unknown. Metabolic acidosis or increased acid load as occurs in chronic renal failure after a prolonged fast or with ethanol ingestion, inhibits the secretion of uric acid.

Hyperuricemia is defined statistically as a serum uric acid level of above 7.5 mg per 100 ml for males and above 6.6 mg per 100 ml for females using the automated colorimetric method of determination. The risk of developing gout correlates with the serum uric acid level. Gout is rare in patients with uric acid levels of less than 7 mg per 100 ml, whereas 83% of patients with a uric acid level greater than 9 mg per 100 ml develop gout. Although the exact reason for the sudden attack of gout in a hyperuricemic patient is unknown, acute attacks may be precipitated by acute fluctuations in serum uric acid level and trauma to the joint. The likelihood of developing gout increases with age.

When urate crystal precipitate in the joint fluid, they are able to stimulate an intense inflammatory reaction. Neutrophils infiltrate the joint space attempting to remove the foreign crystals. During this process, they release bradykinin, proteases, interleukins, and other inflammatory mediators. The clinical result is a swollen, painful, red joint.

The pathognomonic lesion of gout is the tophus, which is sodium urate deposit surrounded by a foreign-body reaction. The water-soluble crystals are anisotropic (negatively birefringent) when viewed under a polarized light microscope. Sodium urate is deposited in cartilage, epiphyseal bone, periarticular structures, and kidneys. Common sites for tophi include the earlobe, the olecranon, and patellar bursas and tendons. Urate deposits in the joints result in cartilage degeneration synovial proliferation and pannus formation, destruction of subchondral bone, proliferation of marginal bone, and fibrous or bony ankylosis.

Sodium urate crystals are found in the medulla of the kidneys with interstitial inflammatory or vascular reaction. The interstitial inflammation, which may be acute or chronic, results in tubular damage.

Acquired hyperuricemia occurs in patients with polycythemia vera, secondary polycythemia, leukemia, lymphoma, multiple myeloma, chronic hemolytic anemia, and after radiation or chemotherapy for a variety of cancers. Both overproduction and undersecretion of uric acid play roles in the development of secondary gout. Serum and urinary levels of uric acid tend to be higher than in primary gout. Drugs that interfere with secretion of uric acid, such as the thiazide diuretics, also may cause secondary gout. Chronic renal disease may cause hyperuricemia, but gouty arthritis usually is not seen. Patients who have had lead intoxication may develop gout due to damage to the kidneys.

The pathophysiology of CPPD is similar to that described for gout excepting that the inflammatory response is not generally as intense.

Symptoms and Signs—Primary gout has three manifestations: asymptomatic hyperuricemia, acute gouty arthritis (which recurs after asymptomatic intervals), and chronic gouty arthritis. For unknown reasons many patients with hyperuricemia never develop gouty arthritis, urolithiasis, or renal damage.

Acute Gouty Arthritis—The onset of the attack is abrupt and typically involves the great toe, although the instep, ankle, or knee may be involved. The pain is intense or excruciating. Fever may be present. The initial attack usually subsides in a few days to a few weeks, and recovery is complete.

The interval following the initial attack may be from a few weeks to many years. Later, the attacks become more frequent, may involve more joints, and are often more severe.

Chronic Gouty Arthritis—Without treatment and after many years, visible tophi develop, permanent joint destruction occurs, and symptoms become chronic. The tophi are relatively painless. However, there is progressive stiffness and persistent aching of affected joints. Destruction of joints and large tophi may lead to grotesque deformities and crippling. The tophi may ulcerate and extrude the chalky sodium urate.

Urolithiasis—Uric acid stones occur in approximately 20% of patients with gout. The development of urolithiasis may precede the acute attack of gout. A predisposing factor to urate renal stone formation is the excretion of acidic urine throughout the day.

Calcium Pyrophosphate Dihydrate Deposition Disease—CPPD is characterized by chondrocalcinosis and acute attacks of pseudogout. The prevalence increases with age. Associations with other diseases such as hemochromatosis, hyperparathyroidism, ochronosis, Wilson's disease, and hypothyroidism have been demonstrated. Pseudogout describes acute inflammatory arthritis in which positively birefringent rhomboid crystals of CPPD are identified on synovial fluid analysis. By far the most commonly involved joint is the knee. Between attacks the joint may be entirely asymptomatic or show changes of osteoarthritis. Radiographic evidence of calcinosis in cartilage and other joint-related structures usually is found.

Hydroxyapatite crystals have been described recently in the synovial fluid of acutely inflamed joints. They are not resolvable by light microscopy and require electron microscopic or microanalytic techniques for identification. The knee and shoulder are most commonly involved.

Diagnosis—The diagnosis of gout requires the examination of affected joint fluid or tophus material under a polarizing microscope, which will reveal the presence of negatively birefringent, yellow, needle-shaped crystals. In CPPD, the crystals are rod-shaped and show a weakly positive or no birefringence by polarizing, compensated microscopy. Patients with CPPD will also commonly have linear densities noted in the articular cartilage of affected joints on x-ray.

Systemic Lupus Erythematosus

This condition is a multisystem disease of unknown etiology that predominately affects young women but can affect men and women of all ages. It often is viewed as the prototypic autoimmune disease in which antibodies are formed against one's own tissues.

Epidemiology—Systemic lupus erythematosus (SLE) is most commonly seen in women between ages 15 and 40, although all ages may be affected. Females predominate over males 5:1. In the United States, Blacks and Hispanics have a higher incidence of disease compared with Caucasians. There is also a strong familial pattern with first-degree relatives of patients having a higher likelihood of disease.

Etiology—Although many potential etiologies, (eg, viral infections) have been proposed, none have been clearly substantiated. A small percentage of patients given procainamide or hydralazine develop a syndrome, which mimics SLE.

Pathophysiology—Antibodies are formed against one's own DNA. These autoantibodies bind the antigen (DNA) and complement, forming immune complexes which, when deposited in various organs, cause injury. The cause for the formation of these antibodies remains unknown.

Symptoms and Signs—The manifestations of SLE are multiple and involve several body systems. The most frequently involved areas are: skin, musculoskeletal, renal, neurological, and hematological.

Skin Manifestations—The most recognized manifestation of SLE is the malar or "butterfly rash" of the face. This is an erythematous elevated rash across the nose and cheeks. SLE may also cause a discoid rash. Discoid lesions begin as erythematous papules or plaques that may scale and become hypopigmented in the center. They may eventually produce scarring. Patients with SLE are frequently photosensitive to sun exposure. Ulcers of the mucous membranes including the mouth and vagina are often seen as well.

Musculoskeletal Manifestations—Arthritis and arthralgias are the most common presenting symptoms and signs of SLE. The arthritis may involve any joint but most often involves the small joints of the hands, wrists, and knees. Generally involvement of the joints is symmetrical. The arthritis is not destructive or erosive, in contrast to rheumatoid arthritis.

Renal Manifestations—Nephritis is suspected by the finding of proteinuria, hematuria, casts, or elevated serum creatinine. The glomerulonephritis may be of several different types. Renal destruction can be rapid and severe in some cases.

Neurological Manifestations—Neuropsychiatric signs and symptoms include seizures, strokes, peripheral neuropathies, cranial neuropathies, intractable headaches, organic brain syndrome, and psychosis.

Hematological Manifestations—Anemia, leukopenia, lymphopenia, and thrombocytopenia are common findings in patients with SLE. The anemia may be due to chronic inflammation, renal disease, or drugs. However, the most significant is a hemolytic anemia due to antibodies directed against red cell antigens. A variety of clotting abnormalities have been described in patients with SLE. The most common of these is lupus anticoagulant. The name is paradoxical as these patients do have a prolonged PTT and yet generally form both venous and arterial clots causing DVT, PE, and arterial thrombosis. Recurrent fetal loss is associated with the presence of lupus anticoagulant.

Other common manifestations found in patients with SLE include serositis meaning inflammation of the serosa of various internal organs. Most often this involves pericarditis, pleurisy, or peritonitis. Splenomegaly and non-specific lymphadenopathy are also frequent findings.

Laboratory abnormalities may include leukopenia, anemia, thrombocytopenia, false-positive serological test for syphilis, abnormal urinary sediment, and proteinuria, antinuclear antibodies (ANA), antibodies against double-stranded DNA, and hypocomplementemia.

Diagnosis—No single test establishes the diagnosis of SLE. Rather, the diagnosis is classically made by the finding of four of eleven possible criteria as established by the American Rheumatologic Association though finding two or three criteria may be sufficient in some cases. The criteria are a combination of many of the above noted physical manifestations and laboratory findings. Most patients with SLE with have a positive ANA.

Scleroderma

Scleroderma is a disease of unknown etiology characterized by fibrosis of the skin and internal organs.

Epidemiology—Scleroderma is a rare disease estimated to have a prevalence of between 19 and 75/100,000. The peak incidence is in the 5th decade of life. Females are affected more commonly than males, and Black females more commonly than Caucasian.

Etiology—The etiology of scleroderma is unknown.

Pathophysiology—This disease is characterized by increased fibrous tissue disposition and obliteration of small vessels in many organs. Lymphocytic and monocytic cell infiltration is frequently seen in the skin early in the disease. Later, the skin is relatively acellular. The vascular lesions are characterized by widespread endothelial abnormalities and an exuberant proliferation of the vascular intima.

Symptoms and Signs—The first symptoms for most patients are Raynaud's phenomenon, swelling or puffiness of the fingers or hands. The skin of hands typically becomes swollen and then firm, thickened, and leathery in appearance. Gradually, this sclerosis of the skin progresses to involve the face and trunk. Fibrosis of the lungs, GI tract, heart, and kidneys is a later finding. Patients also may manifest one or more of a collection of findings referred to as "CREST" (calcinosis, Raynaud's phenomenon, esophageal involvement, sclerodactyly, and telangiectasis). The most-feared complication of scleroderma is malignant hypertension with rapid onset of renal failure.

Diagnosis—The diagnosis of scleroderma is made clinically. There is no laboratory test available to secure the diagnosis.

Polymyositis and Dermatomyositis

Polymyositis (PM) and dermatomyositis (DM) are idiopathic inflammatory diseases of the skeletal muscle. Dermatomyositis also has skin involvement and is often associated with underlying malignancy.

Epidemiology—Both polymyositis and dermatomyositis are rare conditions with an estimated prevalence of 1/100,000 in the general population. Female outnumber males 2:1. The peak incidence is in the 5th decade of life.

Etiology—The etiology of both PM and DM is unknown. However, there is an association with both conditions and malignancy. Patients with PM and DM have been shown to have an underlying malignant tumor 9% and 15% of the time, respectively. Nearly all tumor types can be associated with PM and DM, although there may be a higher incidence of ovarian, lung, pancreatic, gastric, cervical, bladder, and non-Hodgkin lymphoma.

Pathophysiology—The myositis is characterized by both degenerating and regenerating muscle fibers with a mononuclear cell infiltrate. Several antibodies directed at cytoplasmic RNA synthetases, ribonucleoproteins, and other cytoplasmic proteins have been identified. Unfortunately, these can also be found in other autoimmune disorders and none is specific for either PM or DM.

Symptoms and Signs—Muscle weakness is generally the presenting symptom. It is usually slow in onset and gradually progressive. Often patients will have some symptoms for several months before seeking medical attention. The weakness is most often in the proximal muscle groups and symmetrical. Myalgias and muscle tenderness occur in about half of patients. Muscle atrophy is not usually present until late in the disease even when there is severe weakness.

PM and DM may overlap with features of other connective tissue diseases, particularly scleroderma and systemic lupus erythematosus. Raynaud's phenomena, inflammatory arthritis, fever, and weight loss may also be evident.

In DM, there are several characteristic rashes that distinguish it from PM, although the rash may be transient and may have resolved by the time the patient presents with weakness. The most common rash seen is Gottron's sign. This is a violaceous, erythematous, symmetrical rash that occurs on the extensor surfaces of the metacarpophalangeal and interphalangeal joints of the hands. Similar lesions can occur over the elbows and knees.

The heliotrope rash is a reddish-purple rash that occurs on the eyelids and often has associated swelling of the eyelid. Periungual erythema, abnormal nail-bed capillary loops, and cracking of the skin of the tips of the fingers may also be seen.

Diagnosis—The diagnosis is suspected in patients with proximal muscle weakness, elevated muscle enzymes, and abnormal electromyogram. Confirmation is by muscle biopsy showing inflammation and necrosis. The presence of the skin rash distinguishes DM from PM.

Vasculitis

This is a term used to describe inflammatory changes in blood vessels that can lead to necrosis, thrombosis, and obliteration of the involved vessels. Vasculitis can be a manifestation of an underlying systemic disease or constitute the primary process. Understanding of the vasculitides has been hampered by the lack of a universally accepted and clear classification system. Classifications have been based on clinical, histopathological, and etiological considerations. A major obstacle to classification is the fact that vasculitis is a manifestation of several diseases, and most individual cases do not fit precisely into a well-defined category.

Since vasculitis can involve all organs, a multitude of clinical expressions is seen. Many patients have constitutional complaints such as fever, malaise, anorexia, weight loss, myalgias, and arthralgias. Other features include glomerulonephritis, ischemic heart disease, peripheral neuropathy (mononeuritis multiplex) or CNS involvement, pulmonary infiltrates or effusions, ischemic bowel disease, and rash. Laboratory tests usually suggest a nonspecific inflammatory reaction (eg, elevated erythrocyte sedimentation rate). Diagnosis is based on the clinical presentation in conjunction with biopsy and angiographic results.

POLYARTERITIS NODOSA primarily involves medium-sized vessels and is characterized by infiltration of the vessels with polymorphonuclear leukocytes. In a majority of cases, the etiology is unknown but a few patients have hepatitis B antigenemia. The vessel injury may be mediated through deposition of immune complexes of hepatitis B antigen, antibody, and complement, with resultant damage by neutrophils drawn to the lesions by chemotaxis.

HYPERSENSITIVITY ANGIITIS is a small-vessel vasculitis predominantly involving the skin. It appears to be a manifestation of an allergic reaction to an exogenous (drug, infection) or endogeneous (tumor) antigen. The histopathology is described as "leukocytoclastic angiitis," which is vasculitis with neutrophils and their nuclear dust, extravasated red blood cells, and fibrinoid necrosis of the vessel wall.

WEGENER'S GRANULOMATOSIS is characterized by granulomatous vasculitis of the upper respiratory tract (sinusitis, nasal ulcerations, otitis media), lower respiratory tract (cavitary and nodular infiltrates), and glomerulonephritis. There may also be a variable degree of small vessel involvement.

GIANT-CELL ARTERITIS (also called temporal arteritis) is characterized by segmental involvement of large vessels (pri-

marily branches of the carotid artery) with a mononuclear infiltrate including giant cells and destruction of the internal elastic lamina. The most dreaded complication of giant-cell arteritis is sudden blindness due to ischemic optic neuritis.

ENDOCRINOLOGY

Endocrine glands are organs that secrete hormones directly into the blood. The major endocrine glands are the anterior pituitary, posterior pituitary, thyroid, adrenals, parathyroids, pancreas, and ovaries or testes. The anterior pituitary gland controls the function of the other glands with the exception of the parathyroids, pancreas, and posterior pituitary. The hypothalamus and the central nervous system control the pituitary. Hormones regulate metabolism. Endocrine disorders arise when there is excess or a deficiency of a hormone. Most patients with endocrine dysfunction can be treated successfully.

The Hypothalamus

This organ is responsible for the integration of the central nervous system and endocrine system and is particularly related to the physiological response to stress.

Normal Physiology—See Chapter 64.

Anterior Pituitary Disorders

The pituitary gland is located at the base of the brain in the sella turcica. The cells of the anterior lobe secrete the hormones described below.

Normal Physiology—See Chapter 64.

Epidemiology—Tumors of the anterior pituitary account for 6–18% of all brain tumors. The peak incidence is age 40 to 50 years. There is a similar incidence in men and women with the exception of prolactin secreting tumors, which are more common in women.

Etiology—Pituitary adenomas originate from one of the adenohypophyseal cell types of the pituitary gland. Some forms are associated with certain inherited disorders though the majority arise spontaneously.

Pathophysiology—Tumors that cause increased production of TSH, ACTH, GH, and prolactin develop in the anterior pituitary. Only a few tumors that produce increased amounts of gonadotropins have been identified. A tumor that secretes excess TSH is a rare cause of hyperthyroidism. A tumor may secrete excess ACTH and result in Cushing's disease.

Growth hormone-secreting tumors cause gigantism or acromegaly. If the tumor occurs before puberty and closure of the epiphyseal plate, gigantism with generalized overgrowth of the skeleton and soft tissue occurs. After puberty, a GH-secreting tumor causes acromegaly, which is characterized by overgrowth of bone and cartilage in the distal parts of the body such as the face, head, hands, and feet. Acromegaly also is associated with early osteoarthritis, psychological disturbances, glucose intolerance, and hypertension.

Prolactin-secreting tumors, the most common of the functioning pituitary tumors, cause galactorrhea and amenorrhea.

Sheehan's syndrome is destruction of the pituitary due to hypotension during delivery. The clinical manifestations of panhypopituitarism depend on whether the destruction occurs pre- or post puberty. Prepubertal destruction results in stunted growth and lack of sexual development, and may result in thyroid and adrenal insufficiency. Postpubertal destruction results in gonadal, thyroid, and adrenal insufficiency. Large tumors of the pituitary may also lead to generalized destruction and panhypopituitarism.

Symptoms and Signs—Pituitary tumors may cause headaches, loss of temporal visual fields, bilateral hemianopia, loss of visual acuity, and blindness. The other symptoms and signs relate to the excess or lack of hormone(s).

Diagnosis—The diagnosis of these pituitary disorders is made by the determination of serum levels of the respective hormones combined with imaging (usually MRI) of the pituitary gland.

Posterior Pituitary Disorders

The posterior pituitary secretes ADH and oxytocin.

Normal Physiology—See Chapter 64.

DIABETES INSIPIDUS (DI)—Central diabetes insipidus is a disorder due to decreased production of anti-diuretic hormone (ADH) also known as vasopressin. A decrease in the responsiveness of the kidneys to ADH is called nephrogenic diabetes insipidus.

Epidemiology—No specific epidemiological pattern is described.

Etiology—The most common causes of central DI are neurosurgery or trauma, primary tumors (craniopharyngioma, meningioma, lymphoma), metastatic tumor (breast, lung), infiltrative diseases (sarcoid, histiocytosis X), and idiopathic DI. Rarely, central DI is transmitted as an inherited disorder in an autosomal dominant pattern.

Pathophysiology—Without ADH, the kidney is not able to adequately concentrate the urine. This results in increased urine flow and may lead to dehydration. Typically, the thirst mechanism is stimulated and polydipsia ensues. Hypernatremia occurs if the increase in fluid intake is not sufficient to compensate for the loss of free water.

Diagnosis—Central DI must be distinguished from other causes of polyuria and polydipsia. This can be done by using the water restriction test and serially monitoring urine osmolality, volume, and serum sodium concentration.

Symptoms and Signs—The hallmark of diabetes insipidus is polyuria with excessive thirst and polydipsia. In severe forms, the urine volume is 16 to 24 L/day. Micturition may be required every half-hour, day or night. Urine osmolality is low and urine specific gravity is less than 1.005. If intake does not equal output, the patient may become dehydrated severely.

SYNDROME OF INAPPROPRIATE ADH SECRETION (SIADH)—This is caused by continual release of ADH regardless of plasma osmolality.

Epidemiology—SIADH is typically a disease of adults.

Etiology—Increased and unregulated ADH secretion may be caused by CNS disorders such as stroke, hemorrhage, infection, trauma, or psychosis. Several drugs are known to enhance the secretion of ADH or its effect. Drugs commonly associated with SIADH are chlorpropamide, carbamazepine, cyclophosphamide, SSRIs, anti-psychotics, and some chemotheraputic agents. Pain especially following surgery can lead to increased ADH secretion. Lung diseases (pneumonia, Tb, asthma) and HIV also cause SIADH.

Pathophysiology—Increased levels of ADH acts to interfere with the excretion of free water by the kidneys. This leads to progressive dilution of the serum. Hyponatremia is the result of this dilution and can lead to mental status changes, especially in the elderly.

Symptoms and Signs—Ingested fluids are retained, so that volume expansion and dilutional hyponatremia occur. The patient complains of weight gain, weakness, lethargy, and mental confusion. The serum sodium is low, as is plasma osmolality, and the urine is concentrated.

Diagnosis—The combination of hyponatremia, hypo-osmolality, and urine osmolality above 100 mOsmol/kg establishes the diagnosis of SIADH. Generally the urine sodium will be above 40meq/L.

Thyroid Disorders

The thyroid gland, which is located in the anterior neck, secretes thyroid hormones that control a number of metabolic processes.

Normal Physiology—For the biosynthesis and actions of the thyroid hormones, see Chapter 64.

Disorders that affect serum proteins can affect the amount of bound T3 or T4 but not the metabolic status of the patient. Actions of thyroid hormone include maintenance of body temperature and weight, control of skin texture, stimulation of protein catabolism, stimulation of heart rate and myocardial contractility, increased metabolism of cholesterol, and proper functioning of the CNS. At the tissue level the actions of thyroid hormone are synergistic with those of epinephrine.

HYPOTHYROIDISM—This is a state of deficient thyroid hormone production. Cretinism is hypothyroidism that begins at birth and results in developmental abnormalities and severe mental retardation. Myxedema is severe hypothyroidism with the accumulation of hydrophilic mucopolysaccharides in the dermis.

Epidemiology—The prevalence of hypothyroidism is estimated at 0.1–2% of the population. Women are affected 5 to 8 times more often then men.

Etiology—Various mechanisms may cause hypothyroidism. The most common etiology is autoimmune destruction of the thyroid gland (Hashimoto's thyroiditis). Other causes are inherited defects in thyroid hormone synthesis, dietary deficiency of iodine, and disruption of TSH production by the pituitary. The treatment of hyperthyroidism by either surgery or radioactive iodide usually results in hypothyroidism.

Pathophysiology—The lack of thyroid hormone leads to decrease in overall metabolic rate, changes in skin and hair, and affects on some neurological function. The replacement of thyroid hormone generally restores all of these functions.

Symptoms and Signs—The cretin is constipated and somnolent and has a hoarse cry and feeding problems. Physical abnormalities include short stature, coarse features, protruding tongue, broad flat nose, widely set eyes, a protuberant abdomen, and an umbilical hernia. The child is mentally retarded.

Hypothyroidism in adults is insidious in onset. Complaints include cold intolerance, lethargy, constipation, menorrhagia, slowing of intellectual and motor activity, a modest weight gain, dry hair that falls out, dry skin, stiff aching muscles and a deep-hoarse voice. Patients with myxedema have a dull expressionless face, sparse hair, periorbital puffiness, a large tongue, and pale, cool, rough, doughy skin. Coma is a poor prognostic sign.

Physical examination of patients with hypothyroidism is remarkable for the skin changes, bradycardia, and prolonged relaxation phase of deep tendon reflexes. Goiter is caused by hyperplasia of the thyroid gland because of excessive stimulation by TSH in conditions where there is a defect in thyroid hormone synthesis.

Diagnosis—The most sensitive indicator of thyroid function is the TSH level. As thyroid hormone levels fall, the pituitary responds by increasing the production of TSH. Elevated TSH levels are the hallmark of primary hypothyroidism. In some cases, direct measurement of the thyroid hormone level is required.

HYPERTHYROIDISM—This is a state of excess thyroid hormone production and may arise from several different etiologies.

Epidemiology—Graves' disease is the most common form of hyperthyroidism. It is the most common autoimmune disorder with a prevalence of 0.5/1000. Females are 5 to 10 times more likely to have Graves' disease than males. The peak incidence is 40 to 60 years. There is a similar occurrence in Caucasians and Asians, but it is less common in Blacks.

Etiology—There are several causes of hyperthyroidism. The most common, Graves' disease, is caused by autoantibodies to the thyrotropin (TSH) receptor (TSHR-Ab) that activate the receptor, thereby stimulating thyroid hormone synthesis and secretion and thyroid growth (causing a diffuse goiter). Graves' disease is also associated with ophthalmopathy.

Toxic adenoma and toxic multinodular goiter are conditions in which there is focal or diffuse hyperplasia of the thyroid follicular cells, which leads to overproduction of thyroid hormone. Increased iodine load (such as with IV contrast) can lead to hyperthyroidism. Rarely, hyperthyroidism results from a TSH secreting tumor.

Pathophysiology—The excess thyroid hormone leads to increased metabolism, tremor, weight loss, tachycardia, etc. that are characteristic of the disease. In Graves' disease, there is also an increase in the retro-orbital fat that leads to the exophthalmos that is seen.

Symptoms and Signs—Patients with hyperthyroidism may complain of a goiter, a fine tremor particularly when the fingers are spread, increased nervousness, emotional instability, increased sweating, heat intolerance, weight loss, palpitations, weakness, increased appetite, diarrhea, nausea, vomiting, dyspnea, and amenorrhea. Physical examination reveals wasting of muscles, sinus tachycardia, atrial arrhythmias, and perhaps congestive heart failure. The skin is warm, moist, and velvety and the hair, fine and silky. The goiter is usually diffuse, and a bruit may be heard over the gland.

In Graves' disease, the patient also may complain of decreased lacrimation, eye redness, and a sensation of sand in the eyes. The ocular signs include the characteristic stare and frightened facies, infrequent blinking, lid lag, failure of convergence, and failure to wrinkle the brow on upward gaze. Varying degrees of ophthalmoplegia and proptosis occur. Corneal ulceration may occur as a complication. The exophthalmos is usually bilateral.

Diagnosis—The diagnosis of hyperthyroidism is made by detecting depressed levels of TSH and an elevated level of thyroid hormone (except in rare cases of a TSH producing tumor). Radioactive thyroid scans are useful for differentiating hyperthyroidism caused by autoimmune stimulation from autonomously functioning adenoma or goiter. Antibodies to the TSH receptor can also be detected in the serum of affected individuals.

Adrenal Disorders

The adrenal glands produce three principal hormones. Disorders may involve excess or a deficiency of any one or a combination of the hormones. The disorders may be primary, in the adrenal gland, or secondary, due to a problem outside the adrenal gland.

Normal Physiology—See Chapter 64.

CUSHING'S SYNDROME—Cushing's syndrome refers to the presence of excess glucocorticoids.

Epidemiology—ACTH producing adenoma (Cushing's disease) is the most common cause of Cushing's syndrome (other than iatrogenic administration of corticosteroids). There is a female predominance with a female to male ratio of 8:1. The age range is most frequently 20 to 40 years. Ectopic ACTH producing tumors (oat cell carcinoma of the lung) occur 3 times more commonly in males. Cortisol-producing tumors of the adrenal gland occur equally in males and females but are rare with a prevalence of only 2/million in the general population.

Etiology—Cushing's disease is the result of increased cortisol production due to bilateral adrenal hyperplasia caused by an ACTH-producing tumor of the pituitary gland, which acts independently of feedback mechanisms. This accounts for 68% of cases. Nonendocrine tumors, such as bronchogenic carcinoma, bronchial adenoma and pancreatic carcinoma secrete an ACTH-like peptide that causes the syndrome in 15% of cases. Adrenal adenomas or carcinomas are the cause 9% and 8% of the time.

Pathophysiology—Increased levels of glucocorticoids lead to the symptoms and signs that are seen in Cushing's disease and are similar regardless of the underlying mechanism.

Symptoms and Signs—The syndrome is characterized by truncal obesity, hypertension, weakness and fatigability, hirsutism, amenorrhea, purple abdominal striae, edema, and osteoporosis. Approximately 80% of patients have the first four symptoms and signs.

The symptoms and signs of the syndrome are secondary to the excess cortisol. Increased cortisol levels promote the deposition of adipose tissue in the face (the moon facies), in the interscapular area (the buffalo hump), and in the mesenteric bed (the truncal obesity). The obesity is modest, not extreme. Mobilization of protein from peripheral supporting tissue results in muscle weakness, fatigability, osteoporosis, striae, ecchymoses, and easy bruising. Because of increased hepatic gluconeogenesis and insulin resistance, glucose intolerance or diabetes mellitus occurs. Hypertension is almost always present. Marked emotional changes from irritability, emotional instability, and euphoria to severe depression and psychosis occur. Amenorrhea, acne, and hirsutism are seen in females. Acne is seen in both sexes.

Laboratory tests reveal a mild neutrophilic leukocytosis, normal serum sodium, hypokalemia, metabolic alkalosis, and increased serum glucose with intermittent glycosuria. Radiographs show generalized osteoporosis, particularly of the spine and pelvis, and perhaps compression fractures of the vertebrae.

Diagnosis—The diagnosis of the syndrome is based on elevated serum levels of cortisol. Dexamethasone in sufficient doses can suppress ACTH release and subsequently cortisol production in the syndrome due to a pituitary tumor, but will not affect cortisol secretion in the syndrome due to other causes. Patients with an adrenal tumor have increased serum cortisol but decreased serum ACTH.

PRIMARY HYPERALDOSTERONISM—This is due to excessive production of aldosterone independent of angiotensin II.

Epidemiology—The prevalence is estimated to be 0.5% of all hypertensive patients. It occurs in all age groups but most commonly in the 3rd or 4th decades. Conn's syndrome (adrenal adenoma producing aldosterone) occurs twice as often in females than males.

Etiology—Conn's syndrome is primary hyperaldosteronism due to an adrenal adenoma. Approximately 60% of primary aldosteronism is due to an adrenal adenoma, 30% due to bilateral adrenal hyperplasia and the remainder to carcinoma, nodular hyperplasia, or undetermined causes. Secondary hyperaldosteronism occurs in states of overstimulation of the renin-angiotensin system such as in renal vascular hypertension or in hepatic cirrhosis, nephrotic syndrome, or congestive heart failure in which there is a decrease in the intravascular volume.

Pathophysiology—Excess production of mineralocorticoids, primarily aldosterone, leads to sodium retention and hypokalemia due to the effects of these hormones on the kidneys. This results in volume overload and hypertension most commonly.

Symptoms and Signs—The hallmarks of the disease are hypokalemia, hypertension, and volume expansion. The hypokalemia leads to muscle weakness and fatigue, particularly in the legs, and EKG changes. Hypokalemia may predispose to the development of pyelonephritis. The patients complain of polyuria and polydipsia.

Diagnosis—The diagnosis is based on a normal renin levels and elevated urine aldosterone level from a 24-hour collection.

PRIMARY ADRENAL INSUFFICIENCY (ADDISON'S DISEASE)—This is a disease originally described by Addison, which refers to the autoimmune destruction of the adrenal gland with the resultant loss of sufficient cortisol production.

Epidemiology—The prevalence of Addison's disease has been reported to be 39 to 60/million of the general population. The mean age is 40 years.

Etiology—The adrenal glands are destroyed. Approximately 90% of the glands must be destroyed before clinical manifestations occur. Chronic granulomatous infections such as tuberculosis or fungal infection or acute infections such as meningococcemia can cause the destruction. Most cases are due to atrophy of the adrenal glands, which is immunologically mediated and may have a genetic predisposition.

Pathophysiology—Destruction of the adrenal gland is mediated by autoantibodies. Cortisol, which is vitally important for the metabolism of carbohydrates and protein, control of the immune system, and control of vasopressin, is not produced in adequate amounts. Often there is accompanying autoimmune disease of the thyroid or other endocrine glands. In autoimmune adrenalitis, the adrenal medulla,which is the portion of the gland responsible for the production of epinephrine, is usually spared. However, the synthesis of epinephrine depends on high local cortisol concentration. This may lead to inadequate production of epinephrine under physiological stress conditions.

Symptoms and Signs—This disease presents as progressive fatigability, weakness, anorexia, nausea, vomiting, weight loss, increased skin and mucosal pigmentation, and hypotension. Other symptoms include those due to hypoglycemia and abdominal pain, diarrhea, constipation, salt craving, and syncope. Hyperkalemia and hyponatremia due to lack of aldosterone are typically present. The most prominent symptom is fatigue. The hyperpigmentation is brown, tan, or bronze in both exposed and nonexposed areas and particularly over pressure points or in skin creases. The hyperpigmentation results from the over production of ACTH by the pituitary in an effort to stimulate cortisol release from the adrenal gland. ACTH binds to melanocortin-1 receptors in addition to its effects on the adrenal gland and this leads to the hyperpigmentation that is seen clinically.

Diagnosis—A serum cortisol level obtained between 8 am and 9 am is useful to rule out the presence of adrenal insufficiency if the level is >19 μg/dl. Levels <3μg/dl are indicative of the disease. All other patients need dynamic testing. This consists of corticotropin stimulation testing and measurement of serum ACTH levels.

SECONDARY ADRENAL INSUFFICIENCY—This is an ACTH deficiency caused by pituitary destruction or pituitary atrophy secondary to prolonged administration of exogenous corticosteroids. The patient has the same symptoms and signs as the patient with primary adrenal insufficiency (above) but not the hyperpigmentation. ACTH deficiency due to pituitary destruction usually occurs along with other hormone deficiencies. Generally, hypotension is not as problematic because the release of aldosterone is more dependent on Angiotensin II than on ACTH. For this same reason, hyperkalemia is not seen in secondary adrenal insufficiency.

The diagnosis of secondary adrenal insufficiency is made by the finding of low morning cortisol levels and is confirmed by using the insulin-induced hypoglycemia test and following the rise of cortisol in response to hypoglycemia. With secondary adrenal insufficiency, this response will me minimal or absent. A second test known as the short metyrapone test is also available.

ADRENAL CRISIS—This is a state of acute adrenal insufficiency and is life-threatening. Adrenal crisis should be suspected in any patient with unexplained catecholamine-resistant hypotension, especially if they have hyperpigmentation, vitiligo, scanty axillary and pubic hair, hyponatremia, or hyperkalemia.

Etiology—Stress, surgery, trauma, or infection may precipitate acute adrenal insufficiency in a patient who has been chronically adrenally insufficient. Adrenal hemorrhage due to septicemia or anticoagulants or rapid withdrawal of exogenous steroids may precipitate an acute adrenal crisis in a patient with previously normal adrenal function.

Pathophysiology—The symptoms and signs are due to the relative lack of cortisol and catecholamines under conditions of physiological stress.

Symptoms and Signs—The symptoms and signs of chronic adrenal insufficiency become severe and intractable. The nausea, vomiting and abdominal pain are difficult to control and contribute to the dehydration. Somnolence is profound. The blood pressure is low, and the patient may die of hypovolemic shock.

Diagnosis—Finding low serum cortisol levels in a patient with physiological stress makes the diagnosis. A cortisol level of >25 μg/dl in a patient requiring intensive care probably rules out adrenal insufficiency however, a safe cutoff is unknown. A "normal" cortisol level in an acutely ill patient does not exclude this diagnosis, as the level may be normal but insufficient for the physiological state of the patient.

Diabetes Mellitus

This is a disorder of glucose metabolism that results from an absolute or relative lack of insulin and of complications that include accelerated atherosclerosis, retinopathy, nephropathy, and neuropathy. The interrelationship between the glucose intolerance and the vascular disease has not been defined clearly. Type 1 diabetes (insulin-dependent diabetes, formerly called "juvenile onset diabetes") is believed to be an autoimmune disorder. It is characterized by marked insulin deficiency and rapid onset. Late onset and a diminished insulin response characterize type 2 diabetes (also called noninsulin-dependent diabetes, formerly adult onset diabetes).

Epidemiology—This is a disease that occurs worldwide with about 4.2 million diabetics in the US. The incidence is higher in relatives of diabetics, people older than 45 years, and those who are currently or were obese.

Etiology—Both types have a genetic predisposition, which is more obvious in the case of Type 2 diabetes. Destruction of the pancreas by chronic pancreatitis, hemochromatosis, or carcinoma results in diabetes. Other endocrine disorders, such as Cushing's syndrome, hyperpituitarism, and hyperthyroidism, are associated with the disease. Glucose intolerance may occur during pregnancy or times of excessive stress, and at times with the administration of glucocorticosteroids, thiazides, and oral contraceptives.

Pathophysiology—The beta cells of the pancreas are decreased in number or are degranulated in diabetes. The reduction in number of beta cells corresponds to the lack of insulin. In Type 1 diabetes, there are no beta cells. In Type 2 diabetes, only about one-half of them are present. In some cases, these cells are infiltrated with lymphocytes, suggesting an autoimmune mechanism for Type 1 diabetes. The presence of anti-islet antibodies also supports an autoimmune hypothesis in Type 1 diabetes.

The atherosclerosis that occurs in diabetes is the same as the atherosclerosis previously discussed, but it occurs as frequently in females as males and at an earlier age. In the kidneys, nodular glomerulosclerosis (Kimmelstiel-Wilson's) is seen, which is the deposition of glycoprotein in ball-like masses in the mesangial regions of the capillary tufts. Diffuse glomerulosclerosis, which is the deposition of glycoprotein in the mesangium, also is seen, as well as tubular basement membrane thickening. The earliest finding of diabetic retinopathy is microaneurysms. Proliferative retinopathy (the formation of new blood vessels around the optic disk) occurs with long-standing diabetes. Repeated hemorrhages cause scar formation that may lead to retinal detachment. The changes of hypertensive retinopathy also are seen in diabetics with hypertension.

The lack of insulin results in a peripheral underutilization and a hepatic overproduction of glucose, which leads to hyperglycemia. Insulin facilitates the entry of glucose into cells of adipose tissue and muscle, stimulates fat synthesis in cells, and induces protein synthesis. See Chapter 50. The lack of glucose in muscle cells leads to glycogenolysis and the release of amino acids for gluconeogenesis. Lack of insulin and glucose in adipose tissue impairs triglyceride synthesis and promote the release of free fatty acids. The liver metabolizes free fatty acids to ketones, which are used by muscles for energy, to a limited extent. Lack of insulin also results in hepatic overproduction of glucose from glycogenolysis and gluconeogenesis. Another hormone, glucagon, is increased in diabetes. Glucagon effects oppose insulin physiologically.

Hyperglycemia results in glycosuria when the serum level of glucose exceeds the renal threshold for reabsorption of glucose. The osmotic diuresis results in polyuria and polydipsia and may result in dehydration. Excess ketones also are excreted in the urine, as strong acids. This results in urinary loss of bicarbonate and potassium and dehydration.

Normally insulin is released only in response to a glucose load such as a carbohydrate-containing meal. Serum insulin levels rise within 15 to 20 minutes after eating. Patients with Type 1 diabetes do not produce insulin. Those with Type 2 diabetes produce too little insulin and produce it too late to prevent hyperglycemia. Obese people have hypertrophied adipose cells, which, because of their size, are less sensitive to the action of insulin.

The vascular complications of diabetes mellitus have been related to the hyperglycemia. It is postulated that glycoprotein is deposited in the capillaries when glucose levels are elevated. Formation of cataracts and neuropathy are thought to occur because glucose is metabolized to sorbitol by aldose reductase in hyperglycemia. The sorbitol causes osmotic swelling and damage.

Symptoms and Signs—The onset of Type 1 diabetes is sudden and characterized by polyuria, polydipsia, polyphagia, weight loss, decreased muscle strength, irritability, and perhaps a return of bed-wetting. Often the initial presentation may be ketoacidosis. About one-third of these patients have a remission shortly after the onset of the diabetes.

The remission may last for weeks to 1 year, and the patient does not require insulin during this time. After the remission, Type 1 diabetics require insulin for the remainder of their lifetime. They are very sensitive to the effects of insulin and physical activity. Both hypoglycemia and ketoacidosis punctuate their course.

The clinical presentation of Type 2 diabetes may be the insidious onset of weight loss, nocturia, vascular complications, decreased or blurred vision, fatigue, anemia, or symptoms and signs of neuropathy. The disease may be diagnosed from an elevated glucose level without any symptoms. Type 2 diabetics usually are not prone to ketoacidosis. The majority of Type 2 diabetics respond to weight loss.

Diagnosis—The diagnosis of diabetes mellitus is based on the documentation of elevated fasting blood sugar, elevated blood glucose 2 hours after a meal, or an abnormal glucose tolerance test. Diet, physical activity, age, underlying diseases, and drugs influence the accuracy of a glucose tolerance test.

Complications of Diabetes—*Ketoacidosis* occurs in diabetic patients who develop high levels of glucose and ketones plus metabolic acidosis. The usual cause is lack of compliance with insulin therapy but ketoacidosis may be the first episode for an undiagnosed diabetic or a manifestation of an infection. The symptoms and signs of ketoacidosis include nausea, vomiting, abdominal pain, and air hunger (Kussmaul breathing - heavy labored breathing as a compensatory mechanism to the decreased pH). The dehydration may be severe. Oliguria and hypotension may be present. Hyperglycemia, decreased bicarbonate, hypokalemia, azotemia, and acidosis may be seen on laboratory evaluation.

Hyperglycemic hyperosmolar nonketotic coma occurs in Type 2 diabetics. The patients are usually elderly and have some renal impairment. Polyuria and polydipsia precede neurological manifestations. The patient presents with hyperpyrexia, hypotension, tachycardia, hyperventilation, and the signs of dehydration. Hyperreflexia, mild disorientation, confusion, seizures, or coma reflect the intracellular dehydration of the CNS. Laboratory examination is remarkable for increased serum osmolality and hyperglycemia without ketosis or hypernatremia.

Retinopathy occurs in the majority of diabetics after many years of the disease. Venous dilatation, the formation of microaneurysms and small hemorrhages into the retina occur but do not interfere with vision. Hemorrhages into the vitreous cause temporary blindness. Retinal detachment occurs due to repeated hemorrhages and scar formation. Secondary hemorrhagic glaucoma occurs in proliferative retinopathy. Diabetes is the second leading cause of blindness. Cataracts also are associated with the disease.

Neuropathy may result from the sorbitol pathway or from ischemia resulting from the vascular disease. Diabetic neuropathy most frequently involves the peripheral nerves but can involve any nerve. Manifestations of diabetic neuropathy include sexual dysfunction in the male, gastric atony, nocturnal diarrhea, fecal incontinence, orthostatic hypotension, neurogenic bladder, paresthesias, and loss of sensation.

Diabetic ulcers and gangrene result from the neuropathy, the vascular disease, or both. The painless foot is more prone to injury. The ischemic foot is less likely to heal. The patient usually has a history of intermittent claudication, nocturnal leg pain and cramps, loss of hair, and muscle atrophy. Both feet and legs usually become involved.

Nephropathy typically occurs with diabetes of 15 years or more duration and usually occurs along with the other complications. The first sign of diabetic nephropathy is mild proteinuria. Later, the nephrotic syndrome may appear, and renal function deteriorates or progressive renal failure occurs without the nephrotic syndrome. Diabetic nephropathy may cause hypertension. Urinary tract infections and pyelonephritis are more common in the diabetic and may contribute to the renal failure.

Disorders of Calcium Metabolism

These disorders may relate to dysfunction of the parathyroid glands or to vitamin D deficiency.

Normal Physiology—Calcium and phosphate homeostasis is maintained by parathyroid hormone (PTH), vitamin D, and calcitonin. The normal serum calcium varies only slightly for an individual. Dietary vitamin D or that produced in the skin by sunlight is inactive. Vitamin D must be hydroxylated at the 25-position by the liver and at the 1-position by the kidneys to form the active 1,25-dihydroxycholecalciferol. Parathyroid hormone is necessary for the hydroxylation in the kidneys. Parathyroid hormone and vitamin D work together to stimulate gastrointestinal absorption of calcium, bone resorption, and renal reabsorption of calcium. The actions of vitamin D and parathyroid hormone are opposed by calcitonin. Parathyroid hormone promotes the excretion of phosphate by the kidneys. Vitamin D promotes phosphate absorption from the gastrointestinal tract. See also Chapters 64 and 65.

PRIMARY HYPERPARATHYROIDISM—This is an overproduction of PTH with increased serum calcium and decreased serum phosphate.

Epidemiology—Primary hyperparathyroidism is the most common cause of hypercalcemia. It occurs in 0.2% of women over age 65 years and 0.1% of men in that same age group. The majority of cases are sporadic, although there are some with a hereditary cause.

Etiology—Most cases are caused by benign adenomas of one parathyroid gland (80% of cases). Other cases are caused by chief cell hyperplasia in all four parathyroid glands, and a few are caused by carcinoma of the parathyroids. Nonendocrine neoplasms without metastases to the bone that secrete PTH-related peptide cause pseudohyperparathyroidism.

Pathophysiology—The increased level of PTH leads to increased bone resorption, calcium absorption by the gut, and reabsorption in the kidneys resulting in hypercalcemia.

Symptoms and Signs—The majority of patients with primary hyperparathyroidism are asymptomatic, and the diagnosis is discovered after routine screening demonstrates elevated serum calcium.

Some patients present with recurrent nephrolithiasis that leads to urinary tract obstruction, recurrent urinary-tract infections, a predisposition to pyelonephritis, and chronic renal failure. The stones are usually either calcium oxalate or calcium phosphate. Nephrocalcinosis or deposition of calcium in the renal parenchyma also can occur as a result of hyperparathyroidism. Nephrocalcinosis may lead to chronic renal failure. The effect of increased levels of PTH on the bone results in decreased number of trabeculae, increased osteoclasts, and replacement of normal bone by fibrous tissue, which is known as osteitis fibrosa cystica. The hands and skull are affected most commonly. Radiographs show phalangeal resorption.

Increased serum calcium can result in mental status changes from mild personality disturbances to severe psychotic disorders, obtundation, and coma. Proximal muscle weakness, easy fatigability, and muscle atrophy are caused by increased serum calcium. Patients with hyperparathyroidism have a high incidence of duodenal ulcers that may be related to the increased serum calcium.

Other causes of hypercalcemia, such as osteolytic metastases from various malignancies, prostaglandins from various cancers without metastases to the bone, vitamin D intoxication, milk-alkali syndrome, and prolonged immobilization must be excluded. The serum level of PTH is not elevated in these situations.

Diagnosis—Finding an elevated level of PTH in the presence of hypercalcemia makes the diagnosis.

SECONDARY HYPERPARATHYROIDISM—This occurs in situations in which serum calcium levels fall and the parathyroids are intact. Chronic renal failure causes secondary hyperparathyroidism. The failing kidney is not able to hydroxylate vitamin D to the active form resulting in low serum calcium levels. The parathyroid glands secrete more PTH in an effort to stimulate more vitamin D hydroxylation and more bone resorption. Thus, osteitis fibrosa cystica is a part of the bone disease of chronic renal failure. The serum calcium level is normal, although the serum phosphate and PTH levels are high.

HYPOPARATHYROIDISM—The production of PTH is decreased. Pseudohypoparathyroidism is a resistance of the renal tubules to the action of PTH. Serum calcium is low, and serum phosphate and PTH are high.

Etiology—Hypoparathyroidism is caused most commonly by surgical removal or damage to the glands. A congenital absence of PTH occurs rarely. Pseudohypoparathyroidism is an X-linked inherited disorder.

Symptoms and Signs—Hypocalcemia causes neuromuscular irritability, which is manifested by tingling and numbness around the lips, and of the hands and feet. Tetany and convulsions are the most serious manifestations of hypocalcemia.

The patient with pseudohypoparathyroidism is of short stature and has short metacarpals and metatarsals. The serum PTH level is high. In addition to the symptoms and signs of hypocalcemia, these patients may have resorption of bone and soft tissue calcifications as in primary hyperparathyroidism.

OSTEOMALACIA AND RICKETS—This is due to defective mineralization of the normal bone matrix. Osteomalacia refers to the disorder that occurs after the bones have ceased growing; rickets refers to the disorder in growing bones.

Etiology—The defect is a deficiency of vitamin D. Vitamin D deficiency may result from consumption of a deficient diet, inadequate exposure to the sun, intestinal malabsorption of vitamin D (a fat-soluble vi-

tamin), chronic acidosis, renal tubular defects, and therapy with anticonvulsants.

Pathophysiology—A precise concentration of calcium and phosphate is required for mineralization of bone matrix. A deficiency of vitamin D results in decreased absorption of calcium and phosphate from the gastrointestinal tract. Hypocalcemia stimulates the production of PTH, which increases calcium resorption from the bone and phosphate excretion by the kidneys. Mineralization cannot occur because of the decreased calcium and decreased phosphate.

Symptoms and Signs—A child with rickets has skeletal deformities, an increased susceptibility to bone fractures, muscular weakness, hypotonia, delayed dental eruption, defects in the enamel of the teeth, and in severe cases, tetany. Adults with osteomalacia have skeletal pain, bone tenderness, muscular weakness, and fractures of the bones with minimal trauma.

Diagnosis—X-rays of the bones show typical findings. In children, there will be widening of the epiphyseal growth plates, a frayed appearance of the metaphysis, and cupping and widening of the metaphyses. In adults, the radiographic findings are similar to those found in patients with osteoporosis—a generalized loss of bone density with thinning of the cortex.

Serum levels of vitamin D and vitamin D metabolites can assist in the diagnosis.

OSTEOPOROSIS—This is not a disorder of calcium metabolism. The amount of calcium per unit mass of bone is normal in osteoporosis, but the amount of bone is decreased. The condition occurs with aging as bone resorption exceeds bone formation. It occurs in the spine leading to back pain, collapse of vertebrae, and deformity of the spine. Long bones and hips are also susceptible to the disease with subsequent ease of fracture.

The Hyperlipoproteinemias

These result from disturbances in the synthesis or degradation of lipoproteins. The morbidity and mortality associated with this family of disease result from the ability of abnormally high lipoprotein levels to cause atherosclerosis and pancreatitis. Primary lipoproteinemias are due to disorders in lipoprotein metabolism and have a genetic basis, while secondary hyperlipoproteinemias occur because of a concurrent disease such as diabetes mellitus or hypothyroidism. As a complete discussion of all hyperlipoproteinemias is not possible here, only two of the more common primary types, familial hypercholesterolemia and familial hypertriglyceridemia, will be presented.

Normal Physiology—The physiological role of the lipoproteins is to transport lipids (ie, triglycerides, and cholesterol esters) through plasma. Lipoproteins are comprised of triglycerides, cholesterol, phospholipids, and protein (apoprotein). Various lipoproteins differ in the quantity of these components and thus density and size. Lipids are transported in the body by lipoproteins through exogenous and endogenous pathways.

In the exogenous pathway, dietary lipids are incorporated into chylomicrons that are transported to adipose and muscle tissue where the triglycerides are removed. The remainder of the chylomicron, or remnant particle, is transported to the liver for further metabolism.

The endogenous pathway has its base primarily in the liver, where carbohydrates and other substrates are converted to triglycerides. The liver secretes these triglycerides into the blood as very low-density lipoproteins (VLDL). These particles are handled in much the same way as chylomicrons except that after removal of the triglycerides by adipose tissue, a further transformation occurs. Most of the protein is removed, yielding low-density lipoprotein (LDL), which is composed chiefly of cholesterol. These LDL particles supply cholesterol for various uses, including cell-membrane composition and glucocorticoid synthesis. In addition, some LDL particles are degraded by the reticuloendothelial system. As the cells of this system turn over, cholesterol is released and incorporated in the high-density lipoprotein (HDL). Certain components of the HDL apoprotein are transferred to VLDL, and cholesterol is transported back to the liver, a major site of cholesterol synthesis.

FAMILIAL HYPERCHOLESTEROLEMIA—

Epidemiology—This common type affects approximately 1 in 500 individuals in the general population.

Etiology—Familial Hypercholesterolemia (FH) is caused by an autosomal defect in the gene that codes for the LDL receptor. A rare autosomal recessive form is caused by a defect in an adaptor protein for the LDL receptor or for a ligand on the LDL receptor.

Pathophysiology—The defects occurring with this disorder are an inability to bind and/or transport LDL into cells for subsequent catabolism and regulation of cholesterol-synthesizing mechanisms. Thus plasma LDL is elevated. More is taken up by the reticuloendothelial system resulting in accumulations in various locations in the body. These accumulations are called xanthomas. LDL also infiltrates the walls of blood vessels, ultimately resulting in atherosclerosis.

Symptoms and Signs—Patients have high LDL blood levels from birth and throughout life. The chief manifestation is myocardial infarction, which results from coronary atherosclerosis. Myocardial infarctions may occur as early as the 1st decade in homozygotes and generally by the 3rd or 4th decade in heterozygotes. Xanthomas, a common sign of this disorder, increase in frequency with age. They tend to occur in tendons and the eyelids. With the homozygous form of this disease, xanthomas also may form in the skin over the knees, elbows, and buttocks as well as between fingers. High plasma cholesterol (or LDL) yet normal triglyceride levels suggest the diagnosis.

Diagnosis—The diagnosis of FH requires elevated cholesterol levels with usually normal triglycerides and genetic or cellular confirmation of a defect in the LDL receptor. Supportive evidence is the presence of premature coronary artery disease in a first-degree relative or two or more second-degree relatives.

FAMILIAL HYPERTRIGLYCERIDEMIA—This disease involves elevated blood levels of VLDL with resultant hypertriglyceridemia.

Epidemiology—Familial hypertriglyceridemia occurs in approximately 1 in 500 persons.

Etiology—This is caused by an autosomal dominant disorder. It is often associated with obesity, insulin resistance, hyperglycemia, hypertension, and hyperuricemia. The underlying disorder is a mutation in the lipoprotein lipase gene (LPL).

Pathophysiology—The underlying defect is one of several inactivating mutations in the gene for LPL. The incidence of diabetes mellitus and obesity is higher in this patient population and both contribute to the hypertriglyceridemia.

Symptoms and Signs—These patients usually exhibit hyperglycemia, hyperinsulinism, and obesity in addition to hypertriglyceridemia. Such findings usually are not manifested until after puberty. As with familial hypercholesterolemia, atherosclerosis is frequent and may lead to myocardial infarction. Unlike in hypercholesterolemia, xanthomas are not common. In addition to the inherent complications of diabetes and obesity, both contribute to this condition and thus to the atherosclerosis. The diagnosis is suggested by the finding of elevated plasma triglycerides with normal cholesterol levels. Some patients have elevated chylomicron levels in addition to the increased VLDL.

HEMATOLOGY

Normal Physiology-Hematopoiesis—Blood is an organ that performs many functions. It is the transport system for the body. Oxygen, glucose, amino acids, and fats are transported to cells for metabolism. Waste products of metabolism are transported to organs for excretion. Hormones transported by blood regulate the functions of organs and tissues. Blood cells and proteins are responsible for host defenses against infection and cancer. Blood also has the self-preserving function of hemostasis or clot formation. Blood is composed of red blood cells (RBCs), white blood cells (WBCs), platelets, and plasma.

In the embryo, the yolk sac is the blood-forming organ until about 3 months of gestation. The liver and spleen then become the blood-forming organs. These organs do not normally continue to form blood cells after birth. The bone marrow becomes a hematopoietic organ at 6 months of gestation and continues so after birth. An adult has active bone marrow in the axial skeleton whereas hematopoiesis during childhood also occurs in the long bones. With age the bone marrow in the long bones becomes progressively replaced by fat. In disease states where the need for red blood cells (RBCs) is increased greatly, bone marrow may revert to the infant pattern, increasing RBC production 5- to 8-fold. When this compensatory mechanism is exceeded, the spleen and liver may assume some hematopoietic functions. The fetus makes hemoglobin F, which carries oxygen more efficiently at low oxygen tensions. At birth hemoglobin F is replaced largely by hemoglobin A, although production of hemoglobin F continues throughout life, especially in certain diseases. The fetus has a high RBC count, which falls at birth since the increased number of RBCs is no longer needed.

Blood cells follow certain principles of maturation. Bone-marrow stem cells are pluripotent and can become a RBC, WBC, or platelet. During maturation, the size of a blood cell decreases. Young cells are capable of protein synthesis while mature cells, except lymphocytes and

macrophages, are not. The nucleus in a young cell is large and contains loose fine chromatin. A mature cell has a small nucleus without nucleoli and with dense chromatin.

The reticulocyte is the next-to-last step of maturation of the red blood cell. The nucleus is absent in the reticulocyte, but some RNA and ribosomes are still present. These are absent in mature RBCs. Reticulocytes are seen in the peripheral circulation and normally compose 1% of the RBCs. A normal RBC has a life span of 120 days. The production of red blood cells is stimulated by erythropoietin, which is synthesized, in part, in the kidneys in response to hypoxia. Androgens also increase RBC production probably through their effect on erythropoietin.

White blood cells (WBC) are the second component of blood and are primarily responsible for immune functions. The earliest white blood cells evolve along a differentiation line to become several different cell types, each with a distinct function. These include: polymorphonuclear cells (PMNs), lymphocytes, eosinophils, monocytes, and basophiles.

Platelets are small cellular entities that are involved in hemostasis. They are responsible for initiating blood clotting and help form a physical plug at a bleeding site. The precursor cell in the bone marrow is a megakaryocyte.

Plasma contains various proteins. The most abundant of these is albumin. The various coagulation proteins are also found in plasma.

Hematological disorders can affect any or several of these components of blood.

Anemia

Anemia is defined as a decrease in the number of red blood cells. This may occur as a result of RBC loss (bleeding), RBC destruction, or decreased RBC production.

PERNICIOUS ANEMIA—This is a defect in RBC production caused by lack of vitamin B12.

Epidemiology—Pernicious anemia is most commonly seen in people of Northern European decent and African-Americans. It is uncommon in Asian people. The average age of onset is 60 years and is rare under age 30 years. The incidence is substantially increased in patients with autoimmune diseases such as Graves', thyroiditis, vitiligo, adrenal insufficiency, and hypoparathyroidism.

Etiology—Most cases are caused by autoimmune destruction of the parietal cells of the stomach that make a protein known as intrinsic factor which is responsible for permitting absorption of vitamin B12 by the ileum. A total of 90% will have anti-parietal cell antibodies, and 60% have antibodies directed at intrinsic factor as well.

Other causes of vitamin B12 deficiency include total gastrectomy, stomach damage due to corrosives, intestinal malabsorption due to inflammatory disease, resection of the ileum, and competition for vitamin B12 by bacterial overgrowth or the fish tapeworm.

Pathophysiology—It is characterized by lack of intrinsic factor secretion, and consequent atrophy of the gastric mucosa. As a result, vitamin B12, which is needed for proper RBC production, is not absorbed by the ileum of the small intestine. This leads to the decreased production of RBCs by the bone marrow. Those that are produced are characteristically larger than normal RBCs (macrocytic).

Vitamin B12 is also important for normal neurological function. Lack of vitamin B12 leads to demyelination of nerves followed by axonal degeneration.

Symptoms and Signs—The nonspecific symptoms and signs of anemia occur and because of defects in epithelial cells a red, sore, glazed tongue is seen. The neurological abnormalities consist of numbness, tingling, and loss of vibratory sense in the extremities, loss of position sense, loss of fine coordination, spasticity, irritability, memory loss, and mild depression. The GI complaints include anorexia and significant weight loss. Examination of the blood shows oval macrocytes. The red blood cells may be shaped bizarrely (poikilocytosis) and of different sizes (anisocytosis). The reticulocyte count is decreased. The nuclei of the neutrophils may have five or more lobes (hypersegmentation), and there may be mild to moderate neutropenia and thrombocytopenia, with the platelets also bizarre in appearance. The bone marrow shows the megaloblasts, erythroid hyperplasia, abnormal mitoses in the red cell series, large leukocytes with bizarrely shaped nuclei, and decreased numbers of megakaryocytes.

Diagnosis—Macrocytic anemia with low vitamin B12 levels makes the diagnosis.

FOLIC-ACID DEFICIENCY ANEMIA—This is a megaloblastic anemia due to folic-acid deficiency that may be confused with vitamin B12 deficiency anemia.

Etiology—Most cases are due to an inadequate diet. Folic-acid deficiency is seen frequently in alcoholics. A dietary deficiency also may be combined with increased demand, as in pregnancy, hemolytic anemia, hemoglobinopathies, or myelofibrosis. Malabsorption of folic acid occurs in inflammatory small bowel diseases. Certain drugs such as methotrexate, pyrimethamine, triamterene, pentamidine, trimethoprim, and nitrous oxide inhibit conversion of folic acid to its biologically active form. Oral contraceptives, barbiturates, phenytoin, and ethanol have been associated with megaloblastic anemia that responds to treatment with folic acid.

Symptoms and Signs—In addition to the other symptoms and signs of anemia, the patient with folic-acid deficiency may appear wasted. Diarrhea is a prominent complaint. No neurological deficits are attributed to folic-acid deficiency.

ANEMIA OF CHRONIC DISEASE—This is seen in association with a number of chronic inflammatory or infectious diseases.

Pathophysiology—The problem involves a defect that prevents transport of iron from storage depots. The impaired RBC production along with a mildly reduced RBC survival leads to the development of anemia.

Symptoms and Signs—The anemia is usually normocytic normochromic but may be microcytic normochromic or even hypochromic. The serum iron is low, and the total iron binding capacity (TIBC) is normal or low. The saturation index is greater than 10%. The serum ferritin level is normal to high. Increased amounts of iron are stored in the bone-marrow reticuloendothelial system.

Diagnosis—The combination of low serum iron, normal or low TIBC, and an underlying inflammatory or infectious process is adequate for the diagnosis.

ANEMIA OF RENAL FAILURE—This anemia is usually severe and multifactorial in origin.

Pathophysiology—The kidneys are the source of erythropoietin, and production of erythropoietin is decreased in chronic renal failure. The anemia also may be due to iron deficiency because blood is lost from the gastrointestinal and genitourinary tracts in uremia. A hemolytic anemia occurs possibly because of toxins in the blood. Bone marrow is suppressed by the accumulation of toxins.

Symptoms and Signs—Anemia is usually severe, with hematocrit values of 15–30%. However, patients are not as symptomatic as the severity of the anemia would suggest. The anemia is normochromic normocytic unless iron deficiency is also present.

Diagnosis—Typically a low serum erythropoietin level is detected in the presence of renal failure.

HEMOLYTIC ANEMIAS—These involve the destruction of RBCs in the blood stream or by macrophages in the liver and spleen.

Etiology—Hemolysis may be caused by a variety of factors. Antibodies may develop toward RBCs as a result of sensitization, exposure to drugs, infections, or spontaneously. Excessive external trauma, such as marching or jogging, or excessive internal trauma such as occurs with a prosthetic cardiac valve may cause hemolysis. Toxins from the venom of a cobra, the brown recluse spider, and *Clostridium welchii* cause hemolysis. Infections of the RBCs with malaria and bacteremia due to pneumococcus, *Staphylococcus*, and *E coli* cause hemolysis. The RBCs may be made defectively because of an inherited error or have hereditary errors in metabolic enzyme systems.

AUTOIMMUNE HEMOLYTIC ANEMIA—This is characterized by development of IgG or IgM antibodies against the patient's own RBCs.

Etiology and Epidemiology—The disease can occur at any age and may be idiopathic or occur in association with another immune disorder such as lymphoma, chronic lymphocytic leukemia, or systemic lupus erythematosus.

Pathophysiology—The RBCs are coated with an antibody that is directed at one of the many RBC surface antigens. The RBCs are then destroyed in the spleen.

Symptoms and Signs—The anemia ranges from mild to severe. The reticulocyte count is increased. Spherocytes are seen on the peripheral blood smear. Bilirubin is increased. The course is variable but may end in fatal massive hemolysis. The direct Coombs' test is positive. This test uses specific antisera to detect IgG, IgM, or C3 coating the circulating RBCs.

Diagnosis—The diagnosis is based on the evidence of hemolysis (anemia, elevated bilirubin, low haptoglobin) and a positive direct Coombs' test.

DRUG-INDUCED IMMUNE HEMOLYTIC ANEMIA—Three types may occur. Methyldopa induces an autoimmune hemolytic anemia identical to the idiopathic form. The antibody is an IgG against components of the Rh antigen. The direct Coombs' test is positive in about 15% of patients who take methyldopa. There is extravascular hemolysis.

Penicillin and cephalosporins produce a hemolytic anemia by serving as a hapten. The hapten forms a complex with the RBC and antibodies are produced against the drug-red blood cell complex. The hemolysis is extravascular. The direct Coombs' test is positive.

Quinine and quinidine cause hemolysis by the "innocent bystander" mechanism. The drug forms a complex with plasma proteins and IgG and IgM antibodies form against the drug-protein complex. The antibody-drug-plasma protein complex settles on the RBC and fixes complement. C3 remains attached to the RBC. The direct Coombs' test is positive. Intravascular hemolysis occurs. Hemoglobin appears in the urine and acute tubular necrosis may result. See also Chapter 45.

HEMOLYTIC ANEMIA DUE TO HEXOSE MONOPHOSPHATE SHUNT DEFECTS—Glucose metabolism via the hexose monophosphate shunt increases several times when the RBC is exposed to oxidants. The shunt generates glutathione to protect the sulfhydryl group of the hemoglobin from oxidation. Oxidized hemoglobin precipitates in RBCs, forming Heinz bodies. The spleen removes RBCs with Heinz bodies from the circulation. The most common defect in the hexose monophosphate shunt is a hypofunction of glucose 6-phosphodehydrogenase (G6PD) of which there are more than 100 variants. The G6PD gene is located on the X chromosome (sex-linked trait).

Epidemiology—The two most clinically significant forms of G6PD deficiency occur in blacks who originated in Central Africa, and in Eastern Mediterraneans, particularly Sephardic Jews.

Pathophysiology—Some patients with G6PD deficiency are only symptomatic when the RBCs are subject to the stress of infections or oxidants including drugs such as sulfonamides, antimalarials, or nitrofurantoin Heterozygous women have two populations of cells, one with normal enzyme concentration and one deficient.

Symptoms and Signs—Within a few hours of infection or exposure to a drug, the patient has acute hemolysis. Generally, the older RBCs are deficient in G6PD and are destroyed. Therefore, the hemolysis is self-limited even if the exposure to the oxidant continues. The Mediterranean form is characterized by more severe hemolysis. The patient has a decreased hematocrit, increased level of unconjugated bilirubin and hemoglobinuria. A test for G6PD will be falsely negative if done shortly after a hemolytic crisis.

Diagnosis—Deficiency of G6PD can be determined directly by analyzing RBCs in suspected individuals.

SICKLE-CELL ANEMIA—This is the most common conjoint hemolytic anemia. It is due to the substitution of valine for glutamic acid on the β-chain of hemoglobin, which results in hemoglobin S (HbS).

Epidemiology—Approximately 8% of Black Americans are heterozygous or carry the sickle-cell trait. The disease or homozygous form is seen in 0.15% of Black American children.

Etiology—The disorder is inherited according to Mendelian genetics, so that one-fourth of the offspring from heterozygous parents are homozygous, one-fourth are normal and one-half are heterozygous.

Pathophysiology—RBCs must be able to withstand distortion of shape in order to traverse the capillary circulation. RBCs that contain HbS change from biconcave discs to elongated crescent shaped (sickle) cells on deoxygenation. The sickled cells obstruct capillary blood flow, resulting in tissue hypoxia, further deoxygenation of RBCs, and further sickle formation. A small area of ischemia may become a large area of infarction as the process continues. Formation of sickle cells is initially a reversible process, but with time RBC-membrane damage occurs and the sickle formation becomes irreversible. Patients who are homozygous also have 2–20% hemoglobin F, which prevents polymerization of hemoglobin S. RBCs with a high concentration of hemoglobin F do not irreversibly sickle. Any condition that causes hypoxia or dehydration of RBCs increases sickle-cell formation. HbS has decreased affinity for oxygen so the oxygen content of the blood is decreased. Sickled cells are removed from the circulation by the spleen and have an average life span of 15 days.

Symptoms and Signs—Individuals with the sickle-cell trait, but not the disease, usually do not have significant clinical problems. Severe hypoxia is necessary to cause a sickle-cell crisis in these individuals. A person who is homozygous for sickle-cell anemia develops symptoms at about 6 months when much of the hemoglobin F has been replaced. Initial symptoms may be impairment of growth and development. Later, a severe hemolytic anemia develops.

The mortality and morbidity of sickle-cell anemia is related to recurrent episodes of vascular occlusion. A crisis is an episode of sickle-cell formation resulting in severe pain in the chest, abdomen, joints, or other sites. An infection or exposure to cold resulting in vasospasm or conditions that lead to dehydration may precipitate a crisis. The crisis may be mistaken for an "acute abdomen." Chronic organ damage results from recurrent crises. Lung function is decreased because of recurrent pulmonary infarcts. CHF results from the chronic severe anemia, hypoxemia, and pulmonary hypertension. Gallstones develop because of increased bilirubin turnover. Hepatic infarcts may become hepatic abscesses. The hypertonic, hypoxic, acidosis renal medulla is most susceptible to infarction. After repeated infarctions, the ability to concentrate urine is lost. Papillary necrosis also occurs. Prolonged hematuria may result in iron-deficiency anemia. Osteomyelitis may develop in bony infarcts. Aseptic necrosis of the femur occurs. Retinal infarcts, vitreous hemorrhage, and retinal detachment occur. Chronic skin ulcers are seen on lower extremities. Cerebral vascular occlusion can result in stroke, seizures, or coma. With repeated splenic infarcts, splenic function becomes impaired so susceptibility to infection, particularly pneumococcal, increases.

Diagnosis—The diagnosis is made by finding HbS on hemoglobin electrophoresis.

BLOOD DYSCRASIAS

Blood dyscrasia is a term used to indicate a general disorder of the blood. The most common blood dyscrasias include aplastic anemia, agranulocytosis, and thrombocytopenia. Many drugs and chemicals have been cited as the causative agents in blood dyscrasias.

APLASTIC ANEMIA—This term is actually a misnomer. A more accurate description is pancytopenia resulting from damaged pluripotent stem cells. All three cell types—RBCs, WBCs, and platelets—are affected. It is characterized by an acellular or hypocellular bone marrow.

Epidemiology—The overall incidence of aplastic anemia is estimated at 5 to 10 cases per million persons. There are approximately 1000 new cases annually in the United States. Young adults (15–30 years) and the elderly (>60 years) are the most commonly affected.

Etiology—A number of drugs and chemicals have been associated with its production including benzene, chloramphenicol, phenylbutazone, gold, and cancer chemotherapeutic agents. Radiation, infectious hepatitis, and other diseases may also be associated with the condition. Approximately one-half of the cases have no identifiable cause, but recent evidence suggests that many may be due to increased activity of suppressor lymphocytes.

Pathophysiology—The pathogenesis of aplastic anemia is only partially understood. In general, there is one of two pathologic processes that lead to the pancytopenia seen: (1) an acquired intrinsic stem cell defect or (2) an immune-mediated suppression of the bone marrow stem cells.

Symptoms and Signs—The patient complains of progressive weakness and fatigue, mild bleeding from mucous membranes, ecchymoses, and petechiae. The usual signs of infection are not present even though an infection may exist. Symptoms and signs of anemia are present.

Examination of the blood reveals a severe normochromic, normocytic anemia with no reticulocytes. The WBC count is low and is comprised mostly of lymphocytes. There is no increase in bilirubin unless liver disease also is present.

Diagnosis—Bone marrow biopsy is required to make the diagnosis of aplastic anemia.

AGRANULOCYTOSIS—This is characterized by a marked reduction or disappearance of neutrophilic granulocytes in the peripheral blood. Severe neutropenia is defined as less than 500 polymorphonuclear leukocytes (PMNs)/mm^3. The incidence of infection directly correlates with the number of PMNs.

Etiology—Various drugs may cause agranulocytosis, including cancer chemotherapeutic agents, thiouracils, phenothiazines, sulfonamides, or thiazide diuretics.

Pathophysiology—Several mechanisms convey a decreased number of circulating PMNs. Drugs used in cancer chemotherapy as well as radiation will decrease predictably the production of PMNs. This interference with production is usually reversible when the agent is discontinued, unless precursor cells in the bone marrow have been destroyed. Other drugs decrease production of PMNs in an unpredictable fashion and by an unknown mechanism. These drugs include the phenothiazines, sulfonamides, and thiouracils. The decrease in PMNs occurs about 10 days after initiation of therapy with the drug. When the drug is withdrawn the WBC count returns to normal. In some cases the drug may be readministered without problems.

Neutropenia may also result from increased destruction of PMNs. In severe infections the rate of PMN use may exceed the rate of production. Aminopyrine is the prototype for drug-induced granulocytopenia via the "innocent bystander" mechanism. The drug serves as a hapten with plasma proteins, and antibodies are formed against the drug protein complex. The antibody-drug-protein complex settles on the granulocyte and fixes complement. The WBC is removed from the circulation by the spleen. Initially, with increased destruction, production increases but eventually the bone marrow is not able to keep pace.

Symptoms and Signs—The patient may have fever, chills, severe prostration, severe sore throat, and oral ulcers. There is no accumulation of pus at the sites of infection.

Diagnosis—The diagnosis is made by finding an absolute neutrophil count of $<500/mm^3$ on a CBC with differential.

THROMBOCYTOPENIA—A blood dyscrasia characterized by a platelet count of less than $100,000/mm^3$. Spontaneous bleeding may occur when the count is less than $20,000/mm^3$.

Etiology—Thrombocytopenia is caused by one of three mechanisms—decreased production in the bone marrow, increased splenic sequestration, or increased destruction of platelets in the blood stream. There are several causes in each of these categories.

Pathophysiology—*Decreased marrow production*. The most common causes of decreased platelet production are processes that result in infiltration of the marrow. Lymphoma, leukemia, and other tumors can invade the marrow and crowd and reduce the number of megakaryocytes. A number of drugs, such as cancer chemotherapeutic agents, gold, ethanol, thiazides, and sulfonamides, can decrease production of platelets through direct toxic effects on the megakaryocytes.

Splenic sequestration. One third of the platelet mass is normally sequestered in the spleen. Disease states causing splenomegaly such as portal hypertension, splenic infiltration with tumor, or storage diseases such as Gaucher's, cause this percentage of sequestered platelets to increase resulting in lower numbers of circulating platelets.

Accelerated destruction. Destruction of platelets may be immunologic or non-immunologic. The most common immune-mediated causes are viral or bacterial infections, drugs, and idiopathic thrombocytopenic purpura (ITP). Drugs may act as haptens and induce formation of antibodies against the drug-platelet complex. These include quinidine, quinine, analgesics, antibiotics, sedatives, and sulfonamides. Heparin causes thrombocytopenia in approximately 10–15% of patients treated with this agent. The mechanism is generally the formation of a drug-antibody complex binding to the platelet.

Non-immunological mechanisms for platelet destruction may be due to abnormal vessels, fibrin thrombi, and intravascular prostheses. Patients with vasculitis have abnormal vessels that cause platelet destruction. Disseminated intravascular coagulation (DIC), hemolytic-uremic syndrome (HUS), and thrombotic thrombocytopenic purpura (TTP) are all examples of diseases that cause intravascular fibrin thrombi that destroy platelets. Patients with prosthetic heart valves may have low platelet counts from mechanical sheering of the platelets.

Symptoms and Signs—The patient complains of petechiae, purpura, and ecchymoses over the back, upper chest, and limbs and of mucosal bleeding. Blood-filled bullae are found in the mouth. Bleeding may occur from any mucosal surface. Spontaneous bleeding may occur, which may last for several days. The most serious site of bleeding is into the brain. Bleeding time is prolonged.

Diagnosis—Simply finding a low platelet count makes the diagnosis of thrombocytopenia. The underlying cause may require a bone marrow biopsy or antibody titers. Splenomegaly found on physical exam my also provide a clue to the underlying process.

DISORDERS OF HEMOSTASIS

Blood-clotting disorders may result from a defect in any of the steps of coagulation. They may be mild or severe. The coagulation defect may be inherited or acquired.

Normal Physiology—When a blood vessel is cut, two events occur to prevent blood loss—platelet plug formation and blood coagulation. Platelets adhere to the injured vessel surfaces and also aggregate to each other. During adherence and aggregation, platelets assume bizarre shapes with many protruding processes or pseudopodia that overlap.

The next step in hemostasis is blood coagulation. Either the intrinsic or extrinsic coagulation pathway is activated by the surfaces of the injured vessel or by substances liberated by the traumatized tissue or platelets. This process is complete within less than 10 minutes. The clot is composed of a fibrin meshwork with entrapped blood cells, platelets, and serum. The final step in hemostasis is clot retraction, which expresses the serum from the clot and physically draws the torn edges of the blood vessels together. Clot retraction occurs within 1 hour. Clots that form in repairing an injured blood vessel are replaced later by scar tissue. Other clots dissolve.

IDIOPATHIC THROMBOCYTOPENIC PURPURA (ITP)—This usually occurs in young women. An acute idiopathic thrombocytopenic purpura may occur in children following a URI.

Epidemiology—Acute ITP following a viral illness accounts for 90% of cases of thrombocytopenia in children. Most adults with ITP have a more indolent disease that affects women more commonly than men (3:1). Typically it is seen in patients age 20 to 40 years.

Etiology—Acute ITP in children is caused by antibodies directed against viral antigens that cross react with platelet antigens. Adult ITP is an idiopathic autoimmune disorder.

Pathophysiology—IgG, which sensitizes platelets for sequestration by the spleen or liver, develops so that platelet life span is shortened.

Symptoms and Signs—Consists of purpura over the limbs, upper chest and back, and mucosal bleeding. The onset is sudden. No adenopathy, fever, or malaise is associated with the bleeding. The bone marrow shows a normal or increased number of megakaryocytes. The platelet count is low. The bleeding time is prolonged.

Diagnosis—ITP is primarily a diagnosis of exclusion.

HEMOPHILIA A—This is due to an inherited deficiency of Factor VIII activity.

Epidemiology and Etiology—This is a sex-linked recessive trait that occurs in one in 10,000 people. Males and, rarely, homozygous females have the disease.

Pathophysiology—Factor VIII is a large glycoprotein found in trace amounts in normal plasma. It has three components: clot-promoting or antihemophiliac factor activity, antigen, and the von Willebrand factor (VWF). The defect is a deficiency of clot-promoting activity. It may be a defect in the activity of Factor VIII rather then the amount.

Symptoms and Signs—In severe hemophilia, bleeding is often spontaneous, whereas in milder cases excessive bleeding may occur only after injury or surgery. The severity of the bleeding depends on the degree of Factor VIII deficiency. Spontaneous bleeding occurs into joints and muscles. Recurrent hemarthroses are characteristic of the disease and result in permanent joint damage and deformity. Bleeding into the urogenital or gastrointestinal tracts also occurs. Hemorrhage may occur into any organ and may be fatal. Patients with severe hemophilia do not have a normal life span.

Tests of platelet function, bleeding time, and platelet count are normal. The prothrombin time is normal, but the partial thromboplastin time is prolonged.

Diagnosis—An assay finding a low Factor VIII level makes the diagnosis.

HEMOPHILIA B—This is due to an inherited deficiency of Factor IX activity.

Epidemiology—This is a rare sex-linked form of hemophilia.

Pathophysiology and Symptoms and Signs—These are similar to those of hemophilia A.

VON WILLEBRAND'S DISEASE—This is due to an inherited deficiency in von Willebrand Factor (VWF) activity.

Epidemiology—This autosomal dominant condition may be the most common inherited bleeding disorder.

Pathophysiology—VWF is the high molecular-weight component of the Factor VIII complex. VWF supports platelet interaction with the subendothelium. It also carries Factor VIII coagulation activity and prevents its elimination. Defective VWF causes unpaired platelet adhesion and decreased Factor VIII levels.

Symptoms and Signs—Bleeding usually is in mucocutaneous sites. Homozygotes may have symptoms and signs as severe as those in hemophilia. Heterozygotes are often asymptomatic. The bleeding time is prolonged.

Diagnosis—Patients with von Willebrand's disease will have a low Factor VIII level along with an abnormal VWF antigen and ristocetin cofactor assay

VITAMIN K DEFICIENCY—This results in deficiencies of Factors II, VII, IX, and X. Vitamin K is a fat-soluble vitamin found in leafy green vegetables. Stores of vitamin K are limited and deficiency develops in 1 to 3 weeks if intake is stopped.

Etiology and Pathophysiology—Vitamin K deficiency is multifactorial in etiology and involves decreased absorption due to decreased bile acids, impaired intestinal absorption due to inflammatory bowel disease, and changes or decreases in the gut flora, which synthesize vitamin K.

Symptoms and Signs—They are those of bleeding seen in other coagulopathies. The prothrombin time and partial thromboplastin time are prolonged.

Liver disease results in coagulopathy due to decreased synthesis of all factors except Factor VIII. Also, removal by the liver of proteases or enzymes that inactivate the clotting factors is decreased, causing a consumption coagulopathy. The symptoms and signs of the coagulopathy due to liver disease are similar to those of other coagulopathies. The prothrombin time and partial thromboplastin time are prolonged. In addition, hemostasis is further impaired by thrombocytopenia and platelet dysfunction.

Diagnosis—Patients with vitamin K deficiency will have a prolonged prothrombin time (PT), a normal or prolonged activated partial thromboplastin time (PTT), normal bleeding time, and normal platelet count. Confirmation is made by normalization of the PT and PTT following administration of Vitamin K.

DERMATOLOGY

Normal Anatomy and Physiology—The skin is the largest organ in the body. The functions of the skin include sensation, temperature control, prevention of water loss or penetration, synthesis of vitamin D, and protection from organisms and irritants. The skin is composed of three layers: the epidermis, the dermis, and the hypodermic or subcutaneous tissue. The outer layer of the epidermis is the stratum corneum or horny layer. The cells of the stratum corneum are keratinized fully and are without nuclei or granules. In the process of keratinization, the cells from the basal layer migrate upward, flatten, lose water, and fill with keratin. This process normally requires 28 days from formation of a daughter cell (through mitotic division of a cell in the basal layer of the epidermis) until that cell is shed at the surface of the stratum corneum. The cells of the stratum corneum normally are shed invisibly as scales.

The dermis is composed of connective tissue in which are found blood vessels, lymphatics, nerves, arrectores pilorum muscles, fibroblasts, mast cells, and dermal appendages—hair follicles, sebaceous glands, and sweat glands. Elastin and collagen embedded in mucopolysaccharide give the skin its elasticity. Blood vessels in the papillae of the dermis bring nutrients to the avascular epidermis. Sebaceous glands are attached to hair follicles and produce sebum that lubricates the skin and may help prevent water loss. Sebum also has some antiseptic and antifungal properties. Hairs and nails are specialized structures composed of modified types of keratin. The subcutaneous tissues are composed of connective tissue and fat.

Certain microorganisms may be found on the skin as normal flora. Other microorganisms transiently may colonize the skin.

Acne Vulgaris

This is a common disease, which affects teenagers primarily and has, as the characteristic lesions, the open comedo (blackhead) and closed comedo. Most patients have only mild acne and never consult a physician, although they may spend large sums of money on OTC acne aids. In severe forms, acne may lead to extensive scarring. Even the milder forms may cause considerable psychological distress for the patients.

Epidemiology—Acne vulgaris is the most common disorder of the skin in the US. It affects over 17 million patients annually and accounts for 10% of all patient visits to a primary care physician. Almost everyone has some acne during the adolescent years. It may continue in some people until 30 to 40 years of age or appear postmenopausally in women. Administration of certain drugs, such as corticosteroids, halogens, androgens, lithium, and anticonvulsants may result in acne. Acne also may be associated with certain occupations in which tars, oil, and chlorinated hydrocarbons come in contact with the skin. The application of certain cosmetics, including moisturizers, has been associated with acne.

Etiology—The etiology is multifactorial. Heredity plays a role. Androgenic simulation of sebum production by the sebaceous glands at puberty is important, but the main factor in precipitating acne appears to be occlusion of the ducts draining sebaceous glands. There is no scientific evidence that diet commonly plays a role in the development of acne. Anxiety, fatigue, heat, and humidity probably do aggravate acne.

Pathophysiology—The characteristic lesions are the open comedo (blackheads) and closed comedo (whiteheads), which are sebaceous glands that have become plugged with sebum and keratin debris. The black color is he result of oxidation of pigment granules in shed cells forming the plug. When the epidermis covers the opening of the sebaceous gland so that oxidation cannot occur, the lesion is known as a whitehead. Comedones are not inflamed. When they become inflamed, other lesions are formed: papules, pustules, and nodular-cystic lesions. Acne most commonly occurs on the oily areas of the skin, primarily the face, ears, neck, and upper trunk. Healed acne may result in atrophic, pitted, or hypertrophic scars.

Androgens cause sebaceous glands to mature and to produce large quantities of sebum. Both males and females produce androgens. The sebaceous glands respond to very low levels of androgens. Obstruction of flow of sebum from the sebaceous gland to the surface of the skin results in a comedo. Increased amounts of sebum, as well as increased viscosity of sebum and keratin debris, contribute to the obstruction. Chronic obstruction of the sebaceous gland leads to follicular dilatation (enlarged pores). Sebum is composed of triglycerides, waxes, cholesterol, squalene, and minute amounts of free fatty acids. Normally, sebum is not inflammatory. However, bacterial flora in the follicle hydrolyze the triglycerides to free fatty acids, which are extremely irritating and initiate the inflammatory process. In addition, propionibacterium acnes, an anaerobic bacterium that is a normal component of skin flora, thrives on the increased production of sebum. *P acnes* then release chemotactic factors, which enhance the inflammatory process. The inflamed follicle may rupture and spread the process to the adjacent dermis causing increased inflammation via a foreign-body reaction.

Symptoms and Signs—The comedones and other lesions, including scars, are the physical abnormalities of acne. The course is usually chronic throughout adolescence until hormonal balance is achieved, usually in the early 20s. Occasional flares are common during the course. The objective of treatment is to clear the lesions, prevent scarring, and minimize psychological distress.

Diagnosis—Diagnosis is by clinical exam and identification of the classical lesions.

Psoriasis

This is a chronic disease characterized by epidermal hyperplasia and a greatly accelerated rate of epidermal turnover. The lesions are characteristically red, slightly raised, and scaly. Though psoriasis is usually a minor disorder, generalized forms and systemic manifestations also occur.

Epidemiology—Approximately 1–3% of individuals in the US have some form of psoriasis. A higher incidence occurs in Northern European countries, while the disease is rare or absent among Native Americans, Western Africans, Japanese, and Eskimos. Males and females are affected equally. Peak incidence occurs in early and middle adulthood but psoriasis may occur at any time during life.

Etiology—The etiology is unknown. Heredity is thought to play a role, transmission being autosomal dominant with incomplete penetrance or multifactorial. Frequently, the first lesions are associated with previous injury to the site, which is known as the Koebner phenomenon.

Environmental factors such as decreased humidity may aggravate psoriasis.

Pathophysiology—The histopathological changes include parakeratosis (retention of nuclei in cells in the keratin layer), hyperkeratosis (increased thickness of the keratin layer), hypogranulosis (loss of the granular layer), elongation of the epidermal rete ridges, pustules with surrounding intercellular edema (spongiform pustules), and papillomatosis (increased height of the dermal papillary pegs) with thinning of the suprapapillary epidermis. There is an inflammatory infiltrate in the upper dermis and proliferation of small blood vessels in the papillae (vascular ectasia). Mitotic figures are seen in the bottom three cell layers of the epidermis rather than just in the basal layer.

The characteristic change is the markedly shortened rate of turnover and accelerated production of the epidermal cells. Instead of the normal 28 days from cell division in the basal layers until the cell is shed from the stratum corneum, in psoriasis it takes only 3 to 4 days for this to occur. The mechanism for this and the other symptoms and signs of psoriasis is not understood at this time.

Symptoms and Signs—The lesions of psoriasis are discrete or confluent erythematous plaques and papules covered with white or silvery scales. The lesions are characteristically found on the extensor surfaces such as the elbows and knees and also on the back and scalp. However, any area of skin can be involved. Nails commonly are involved with pitting and ridging, while mucous membranes rarely are involved. The lesions may be localized or generalized and are usually asymptomatic but may cause discomfort from burning and itching. Auspitz sign is characteristic (punctate bleeding that occurs when psoriatic scales are removed). The onset of psoriasis is usually insidious, although it may be explosive. The clinical course is chronic and recurring with periods of remission. Spontaneous cures rarely occur. Most cases are only cosmetically disfiguring. Some forms such as psoriatic erythroderma and pustular psoriasis may be life threatening. Although pustular psoriasis

looks like an infection, the lesions are sterile. A form of arthritis that resembles rheumatoid arthritis, but affects the distal joints is associated with psoriasis in some cases. There are no characteristic laboratory abnormalities of psoriasis.

Diagnosis—Diagnosis is by clinical exam and identification of the classical lesions.

Allergic Skin Diseases

URTICARIA (HIVES)—Urticaria is a skin reaction composed of transient wheals (edematous papules and plaques, usually pruritic). Urticaria may be acute or chronic. Immediate reactions occur within 1 to 60 minutes of exposure to the antigen and are manifested by generalized pruritus and urticaria. IgE is the mediator of immediate reactions. These reactions are the most dangerous, since they may be associated with laryngeal edema and/or anaphylaxis. Accelerated reactions occur within 1 to 72 hours of contact with the antigen and also are manifested by generalized urticaria and pruritus. An exanthem is seen rarely with this type of reaction. A late reaction may occur from 3 to 21 days after exposure to the antigen. The urticaria in this case may subside even though the exposure to the antigen is not terminated because of the development of IgG and IgA blocking antibodies. Chronic urticaria lasts longer than 30 days and is rarely IgE mediated. The etiology is unknown in 80–90% of the cases. Emotional stress often exacerbates this condition.

Epidemiology—Approximately 15–23% of the population will experience at least one episode of urticaria. Young adults are afflicted most frequently by the acute form. The chronic form, lasting longer than 4 weeks, usually is seen in patients over 35 years. Individuals with urticaria, or their family members, are likely to be allergic to a number of antigens and also suffer from seasonal rhinitis, asthma, and atopic dermatitis.

Etiology—Urticaria can be caused by IgE or complement mediated reactions usually in response to an antigen. Common antigens are food (milk, eggs, shellfish, nuts, wheat), drugs (penicillin), and parasites.

Urticaria is also caused in some individuals by physical stimuli. Cold urticaria occurs most often in children or young adults. These individuals develop urticarial lesions when the skin contacts extremely cold stimuli such as ice. Similarly, exposure to sunlight may cause urticaria in others. Urticaria may also result from exercise to the point of sweating, prolonged pressure on the skin, or vibration.

Hereditary angioedema is a severe form of recurrent, episodic urticaria that is inherited in an autosomal dominant pattern. This form is characterized by low levels of C1 esterase inhibitor.

Urticaria has also been associated with some connective tissue disorders such as SLE or Sjögren's syndrome. Often this is a sign of underlying urticarial vasculitis. Some bacterial infections, and underlying occult malignancy may also cause urticaria.

Pathophysiology—Urticaria may develop as a result of several different processes, although all involve liberation of histamine from mast cells in the dermis. Systemic exposure to an antigen may result in formation of IgE antibodies toward that antigen. The antibodies are fixed to mast cells in the dermis and lungs and to circulating basophiles. The interaction of antigen and antibody results in liberation of histamine and other mediators (prostaglandin E and kinins). These substances cause arteriolar dilatation and increased capillary permeability in the skin. Histamine is degraded quickly in tissues so urticaria seldom lasts for more than 48 hours. A degranulated mast cell is refractory to further stimulation until histamine granules reform.

Other antibodies may be involved in the liberation of histamine and mediator substances from mast cells. IgG and IgM may be formed against antigens. These antibodies, when they interact with antigens, may activate the complement cascade, which results in histamine release. Cold and solar urticaria are mediated by antibodies (IgE) that are only active at decreased temperature or upon exposure to light.

Histamine may be released from mast cells by nonimmunological mechanisms. Certain chemicals stimulate mast cells directly to liberate histamine. These chemicals include drugs such as morphine, codeine, dextrans, and crayfish toxin or snake venom. Direct physical pressure may also cause release of histamine from mast cells.

Symptoms and Signs—Urticarial lesions are well-circumscribed discrete wheals with erythematous raised serpiginous borders and blanched centers. The lesions, which involve only the superficial layer of the skin may be scattered, localized, or may coalesce. The patient will complain of intense pruritus or burning. Urticaria alone is seldom life threatening, but it may indicate a future anaphylactic reaction. Skin testing is usually of little value in these individuals in that they are allergic to numerous antigens. The acute form usually lasts less than 6 weeks. The chronic form may last for years but usually does not last for life.

Diagnosis—Diagnosis is by clinical exam and identification of the classical lesions.

ATOPIC DERMATITIS (ECZEMA)—Eczema is chronic pruritic inflammatory skin disease of the epidermis and dermis. It is usually associated with a personal or family history of hayfever, asthma, allergic rhinitis, or atopic dermatitis. Atopic dermatitis is characterized by itching. The appearance and distribution of the lesions depends on the age of onset.

Epidemiology—The onset is typically early in life, often in the first 2 months and by 1 year of age in 60% of cases. There is a slight male predominance. Over two thirds of patients have a personal or family history of allergic rhinitis, hay fever, or asthma.

Etiology—The etiology is unknown. Those with onset in early childhood tend to improve after a period of time. Irritants, excessive bathing, wide temperature variation, low humidity, and nervous tension may aggravate the condition.

Pathophysiology—Pathological changes are those of nonspecific dermatitis. Epidermal vesicles due to intercellular edema, parakeratosis, acanthosis, and an inflammatory infiltrate of the epidermis and dermis are seen in acute atopic dermatitis. In the chronic form, hyperkeratosis, parakeratosis, acanthosis, and a lymphocytic infiltrate of the thickened upper dermis are seen.

The mechanisms for the development are not understood. Various immunological hypotheses have been put forth to explain the development of atopic dermatitis, but none of these explains all cases. Some patients with the disease have elevated levels of IgE and perhaps elevated levels of IgG and IgM. It also occurs commonly in immune-deficient individuals and may be due to an impairment of delayed hypersensitivity or impaired phagocytosis. Depressed IgA also has been reported in atopic patients.

Symptoms and Signs—Infant-type atopic dermatitis (infantile eczema) begins during the first few months of life, perhaps as a reaction to food, although this is controversial. The eruption may be local or generalized, acutely inflamed, vesicular or popular, and spreads rapidly. The scalp, face, trunk, extremities, and diaper area are frequently involved. There is considerable oozing and crusting associated with the lesions along with intense pruritus. The skin may become infected secondarily. The child usually outgrows the disease spontaneously at 2 to 3 years.

Childhood-type atopic dermatitis may be a recurrence of infant type or may be the first appearance of the disease. In contrast to the vesicles and oozing of the infant type, these lesions are dried, lichenified plaques and patches. The lesions also are more localized in the childhood type and are found on the flexor surfaces and the face, neck, feet, genitalia, and scalp. Again, there is intense pruritus. The disease may clear or persist into adulthood.

In adult type atopic dermatitis, the lesions consist of chronic lichenified patches, which are intensely pruritic and may be hyperpigmented. Commonly flexures and the creases of the neck and eyelids are involved as well as the same areas as in the childhood type. The clinical course is chronic and characterized by spontaneous exacerbations and remissions. Eventually, the disease fades.

ALLERGIC CONTACT DERMATITIS—This is an extremely common skin disease caused by direct contact with the substance and the development of delayed hypersensitivity. Primary irritant contact dermatitis is caused by contact with noxious agents such as acids or corrosives or excessive contact with soap and water. The inflammatory skin reaction that results from such a contact occur in all individuals exposed to these agents and does not involve the development of hypersensitivity.

Epidemiology—Many patients have this disease, which affects any age group and is equally common in both sexes.

Etiology—Substances capable of forming a stable bond with cutaneous proteins and being transported to a lymph node are allergens for contact dermatitis. These include Rhus (poison ivy and poison oak), ragweed preservatives, solvents, rubber, low-molecular-weight polymers, metals, particularly nickel, and medications.

Pathophysiology—The chemical group binds to skin protein and is transported to the lymph nodes. Cellular proliferation occurs in the paracortical area of the lymph nodes. Small lymphocytes become sensitized to the antigen within 7 to 10 days of the first exposure. The sensi-

tized lymphocytes react with the antigen and release soluble chemotactic factors, which attract other lymphocytes and macrophages into the area. Also, the sensitized lymphocytes release migratory inhibitory factors that inhibit the movement of macrophages and other cells away from the area lysosomal enzymes released from the macrophages result in skin destruction. On subsequent exposures, reaction will occur within 24 to 48 hours.

Symptoms and Signs—The distribution of the lesions is characteristic the rash occurs where the allergen comes in contact with the skin. The scalp is rarely involved. The lesions begin as intense, relatively limited areas of erythema that are soon associated with edema. Papules and vesicles form, with subsequent oozing and weeping. Sometimes the lesions are bullous. The erythema lessens and is replaced with crusting and scaling. Pruritus in varying degrees of severity is always present. If contact with the allergen is eliminated, healing occurs in 1 to 3 weeks. With chronic exposure, a chronic contact dermatitis may develop with thickening, fissuring, scaling, and hyperpigmentation of the area. Vesiculation is minimal in the chronic form. Intense itching and burning may result in excoriation and secondary infection. The disease will recur if there is another contact with the allergen

Diagnosis—Diagnosis may be made via patch testing, although the patient may react to a variety of allergens including some, which he/she is not allergic to.

PHOTOALLERGIC REACTION—Uncommon delayed hypersensitivity reactions that require three factors: light, skin, and an allergen. Distribution is limited to the areas exposed to light. Photoallergic reactions must be distinguished from the more common phototoxic reaction, which occurs when a photosensitizing substance ingested or applied externally, plus minimal exposure to sunlight or artificial lighting, results in an exaggerated sunburn in 6 to 18 hours. No immunologic mechanisms are involved in phototoxic reactions, which can occur with the first exposure to the substance. Pigment is protective in the phototoxic reaction and tanning results as the reaction subsides.

Epidemiology—These reactions are rare but occur predominantly in males (7:1) and in the age group of 40 to 60 years. Pigment and dark skin are not protective for this reaction.

Etiology—Numerous drugs, chemicals, and cosmetics can cause both phototoxic and photoallergic reactions.

Pathophysiology—The energy of light depends on the wavelength in the electromagnetic spectrum. A molecule, when exposed to light, may dissipate the absorbed energy as heat or may undergo one of numerous photochemical reactions including chemical-bond formation. The chemical and the cutaneous protein are the antigen for the development of delayed hypersensitivity

Symptoms and Signs—A photoallergic reaction occurs as an urticarial or eczematous eruption in the areas of sun exposure. The initial eruption will not be seen until 7 to 10 days after the first exposure but occurs within 24 to 48 hours on subsequent exposures. No tanning occurs as the reaction subsides. The reaction may recur with each re-exposure. Photopatch testing may make the diagnosis.

Adverse Reactions to Drugs as Manifested by the Skin

Cutaneous reactions are among the most common adverse reactions to drugs. The significance of these reactions varies from minor to life threatening. Nonallergic drug reactions of the skin include alopecia, purpura, secondary infections, and phototoxic reactions. Allergic reactions include urticaria, the rash seen with serum sickness, allergic contact dermatitis, and photoallergic reactions as already discussed. In addition, several less-common but potentially more serious reactions may occur.

EXFOLIATIVE ERYTHRODERMA SYNDROME—Exfoliative erythroderma syndrome (EES) is a serious, and at times life-threatening reaction of the skin characterized by generalized erythema and scaling associated with fever and generalized lymphadenopathy. It may be a reaction to a drug or may be an extension of a preexisting skin disorder.

Epidemiology—Exfolative erythroderma syndrome is almost always seen in patients over age 50 years and is more common in males.

Etiology—EES is seen as a generalized spreading of a drug reaction, psoriasis, contact dermatitis, seborrheic dermatitis, atopic dermatitis, or in association with leukemia or lymphoma. In 10–20% of patients no underlying cause can be identified.

Pathophysiology—The pathophysiology is entirely unknown.

Symptoms and Signs—There is a generalized erythematous eruption with scaling involving the entire skin surface. In extensive exfoliative erythroderma syndrome, the metabolic demand is such that the patient develops negative nitrogen balance, edema, hypoalbuminemia, and loses muscle mass. Serious water and electrolyte imbalance can result from the greatly increased loss of water through the skin. The cause and complications determine the course. The erythroderma syndrome persists in patients with malignancy. If psoriasis, atopic dermatitis, or other skin diseases cause EES, improvement occurs over 8 to 10 months. Prognosis is better if the etiological factor can be removed. Approximately 30% of patients with EES die.

Diagnosis—The finding of generalized skin erythema and scaling make the diagnosis.

ERYTHEMA MULTIFORME—This is a characteristic skin disorder that occurs as a result of a systemic allergic reaction to various agents. The syndrome may include only a few typical skin lesions or become a more severe illness known as Stevens-Johnson syndrome.

Epidemiology—Erythema multiforme most often affects patients age 20 to 30 years, although 50% of cases are in patients under age 20 years.

Etiology—Infectious agents including herpes virus and *Mycoplasma* pneumoniae, drugs (especially penicillin, aspirin, phenytoin, allopurinol, and sulfonamides), and malignancy may cause this condition. More than 50% of cases are idiopathic.

Pathophysiology—Histopathologically, the changes seen are those of spongiotic dermatitis with epidermal necrosis, ballooning, and vacuolar alteration. An associated superficial perivasculitis and interface lymphohistiocytic infiltrate is present. This disease is probably antigen-antibody-mediated.

Symptoms and Signs—The lesions may be papules, macules, urticaria, vesicles, or bullae. The type of lesion may change as the disease progresses. The lesions are symmetrical in distribution and are found most commonly on extensor surfaces, the backs and palms of hands, and the tops and soles of feet. Both mucous membranes and skin are involved in the severe form. The lesions begin as a bright redness that extends peripherally as the center pales, becomes firm, and may contain bullae. These classical lesions are called target lesions because of their appearance, but do not always occur in the disease. In Stevens-Johnson syndrome, the skin, conjunctiva, and mucous membranes are involved. This reaction includes toxemia, prostration, high fever, cough, and inflammation of the lungs. The disease usually resolves within a few weeks after the inciting agent is removed although the severe form may be fatal.

Diagnosis—The diagnosis is made by the finding of the typical skin lesions on physical exam and supported by skin biopsy.

Skin Infections

IMPETIGO—This is a common superficial bacterial infection of the skin that may arise in normal skin or as a secondary infection of dermatitis, intertrigo, infestations, other infections, or trauma.

Epidemiology—Impetigo occurs primarily in children and young adults.

Etiology—The causative organisms are beta hemolytic streptococci and coagulase-positive staphylococci. In secondary forms, gram-negative organisms also may be found. The lesions may be autoinoculable and are somewhat contagious.

Pathophysiology—Impetigo is caused by the invasion of the superficial layers of the skin by the offending organism.

Symptoms and Signs—The disease begins as a macule that progresses to a vesicle covering about 2 to 3 cm^2 in area. The vesicle, which is located just below the stratum corneum, becomes a pustule filled with polymorphonuclear leukocytes. The pustule ruptures and may spread the bacteria to the adjacent skin. The lesion then becomes denuded and seeps. The seropurulent fluid quickly dries, forming the characteristic friable honey-colored crust of the disease.

Diagnosis—The diagnosis is made by the identification on physical exam of the typical lesions and may be confirmed by Gram's stain and culture.

Mycotic Infections

Dermatophytoses are mycotic infections of the skin that involve the epidermis, nails, and hair. The diseases differ as to causative organism, area affected, mode of transmission, and response to therapy. These infections may occur in any age group.

Tinea capitis usually occurs in prepubertal children and may occur in epidemics in schools or institutions. The lesions are found on the scalp and appear as scaly, crusted patches with the hair broken off close to the scalp. Inflammation and deeper lesions may occur and may result in scarring alopecia. The fungus is of the Microsporum genus.

Tinea corporis is classic ringworm. The lesions occur anywhere on the glabrous skin of the body. A papule begins and spreads centrifugally as a scaly red rim with central clearing. The border of the lesions may contain vesicles. The causative organisms are of the *Microsporum* and *Trichophyton genera*.

Tinea cruris is known more commonly as "jock itch." The lesions begin as a symmetrical scaly red eruption of the groin and inner thighs. Chronic lesions are browner in color. The lesions have specific margins and the margins are more inflamed than the center. Severe pruritus accompanies the eruption. The fungus belongs to either *Epidermophyton* or *Trichophyton* genus. Heat and humidity are aggravating factors for the development of Tinea cruris. This condition must be distinguished from a similar eruption that is caused by another fungus, *Candida albicans*.

Tinea pedis (athlete's foot) is perhaps the most common of the dermatophytoses. Darkness, heat, and humidity predispose an individual to the development of this infection. *Trichophyton mentagrophytes* causes an inflammatory eruption with vesicles and weeping. *Trichophyton rubrum* causes a dry, scaly eruption.

Tinea unguum is a fungal infection of the nails; most commonly the toe nails. The nails become yellow in color, brittle, thickened and raised by the underlying debris. Infections of the nails are difficult to eradicate.

Diagnosis—The diagnosis of any of the mycotic skin infections depends on the physical findings during examination. It is confirmed by skin scrapings and microscopic examination revealing the invading fungus.

INFECTIOUS DISEASES

Urinary Tract Infections

Urinary tract infection (UTI) refers to bacteria multiplying in the urinary tract. It is the most common bacterial infection seen in the US. UTIs are broadly divided into complicated and uncomplicated UTIs.

An uncomplicated UTI, exemplified by cystitis, is a UTI in an anatomically normal urinary tract. Complicated UTIs are those infections of the urinary tract that are associated with a condition that increases the risk of treatment failure. These conditions may include anatomical abnormalities of the urinary tract or the presence of a catheter.

Acute pyelonephritis is a bacterial infection of the kidney. It primarily arises under one of two circumstances. First, if there is vesicoureteral reflux of infected urine. This is a potential long-term problem that can result in recurrent episodes of infection and is due to an anatomical abnormality. Second, a normal urinary tract may become infected with an uropathogenic strain of *E coli* whose *P fimbriae* permit ascent of the urethra without being washed out. This demands adequate immediate treatment but is not a long-term hazard.

Prostatitis and urethritis are infections of the prostate gland and urethra, respectively. Prostatitis often requires a longer course of therapy for complete eradication of the infection.

Each of these infections may be asymptomatic but each has characteristic symptoms and signs.

All urinary tract infections may be either acute or chronic.

Normal Anatomy and Physiology—The urinary tract is a closed system for drainage of urine from kidneys to bladder and eventually to the outside via the urethra. Under normal circumstances the entire urinary tract except for the anterior urethra is sterile. Various defense mechanisms prevent infection in the urinary tract.

The outward flow of urine serves to wash out organisms. This is probably the most important defensive mechanism and can clear 99% of organisms experimentally inoculated into the bladder. Urinary tract anatomy prevents retrograde flow of urine. Valves at the ureterovesical junction prevent reflux of urine from bladder into the ureters and thence the kidneys. Females have a shorter urethra than males (4 cm versus 12 cm), which contributes to the much higher incidence of urinary tract infections in women. Also organisms from the adjacent vagina or rectum colonize the urethra in women easily.

The urine itself has certain characteristics that discourage bacterial growth. These include an acidic pH (5.5), as bacteria prefer a more alkaline medium (pH = 6 to 8); low osmolality, usually below that required for optimal bacterial growth; and the presence of urea and weak organic acids. Prostatic secretions also are probably antibacterial.

The kidney is particularly susceptible to infection because of the hypertonic state of the papillae and medulla. This leads to impairment of leukocyte migration, complement activity, and phagocytosis, as well as development of spheroplasts or protoplasts by bacteria, which make them less susceptible to antibiotics.

Epidemiology—The incidence of urinary tract infections depends on the age, sex, sexual activity, and underlying diseases in the population. Women have a 10–20% lifetime risk of a UTI. The annual incidence is around 1% until adolescence and rises to 10% by age 50. Incidence is much lower in celibate women and higher during pregnancy. A total of 20% of pregnant women with bacteriuria develop acute pyelonephritis. In infancy, the rate of UTIs in males, usually associated with a structural anomaly, exceeds that of females. Urinary obstruction from an enlarged prostate accounts for the rate in elderly men being even higher than that in women. Men under 50 years rarely have UTI unless they are uncircumcised, have an anatomic abnormality of the urinary tract, engage in unprotected insertive anal intercourse, or have AIDS with a CD4 T cell count under 200/μl. Long-term indwelling catheters facilitate entry of uropathogens and hinder their clearance.

Etiology—Most UTIs are caused by gram-negative organisms that normally inhabit the large intestine. *Escherichia coli* accounts for 85% of first urinary tract infections. Other organisms, including *Klebsiella*, *Enterobacter*, *Proteus*, and *Pseudomonas*, are seen less commonly. Instrumentation of the urinary tract is a predisposing factor for development of an infection, particularly with *Proteus* or *Pseudomonas*. *Neisseria gonorrhoeae*, *Chlamydia*, and vaginal organisms may cause urethritis.

Pathophysiology—Bacteria that ascend the urinary tract through the urethra cause most urinary tract infections. This ascent is easier in the shorter urethra of females. The anterior urethra is normally colonized by bacteria from the large intestine in females. The trauma to the female urethra that occurs during sexual intercourse can result in the entrance of bacteria in to the bladder. Instrumentation of the lower urinary frequently results in infection. Bacteriuria commonly occurs within 24 to 48 hours after the placement of an indwelling urinary catheter. The rate of acquisition of catheter-associated bacteriuria is 2–6% per day for each day of catheterization.

Normally, flow of urine washes out any bacteria that enter the bladder. However, certain conditions interfere with this flow and therefore predispose the individual to the development of UTIs. Tumors, stones, strictures, bladder diverticulum, anatomical abnormalities, and prostatic hypertrophy may impede flow of urine. Structural abnormalities, as well as a neurogenic bladder, may prevent complete emptying of the bladder and allow bacteria to remain and multiply in the residual urine.

Conditions that allow retrograde flow of urine increase the incidence of infection. In vesicoureteral reflux, urine from the bladder is forced up the ureters and perhaps into the renal parenchyma by increased pressure in the bladder during voiding. Urethrovesical reflux may draw contaminated urine back into the bladder from the urethra during coughing, sneezing, or laughing. In pregnancy the urine flow is obstructed partially by the enlarged uterus. This results in dilation of the ureters and decreased peristaltic activity of the bladder, allowing for reflux.

Certain uropathogenic strains of *E coli* possess adhesins that bind to receptors on the surface of urinary epithelium. These adhesins allow *E coli* to resist being "washed out" from the urinary tract. The best known form of adhesion is by *P fimbriae* on the bacterial cell wall. *P fimbriae* attach to the carbohydrate moiety of a glycolipid in the epithelial cell. It is adhesion to this receptor that is apparently interrupted by substances in cranberry juice.

Rarely, UTIs may be caused by the hematogenous spread of bacteria from other sites. This usually involves seeding of the kidney by staphylococci.

Symptoms and Signs—Urethritis is accompanied by symptoms related to micturition, including urgency and dysuria. Cystitis is characterized by symptoms of frequency, urgency, dysuria, and perhaps pain or pressure in the lower abdomen. Systemic symptoms or signs are uncommon with cystitis. Acute pyelonephritis is manifested by symptoms

that develop over a few hours to 2 days, including aching pain in the lumbar region (flank pain), fever to 39°, shaking chills, nausea, vomiting, and local symptoms of urgency and dysuria. On physical examination, there may be tenderness over the kidney in the area of the costovertebral angle.

The urinalysis in a UTI may show bacteria, leukocytes, red blood cells, and epithelial debris. White blood cell casts indicate pyelonephritis.

In patients with acute urinary tract infections, the symptoms may resolve with or without therapy. Acute pyelonephritis may resolve spontaneously or recur over many years. Patients without underlying disease usually do not have continuing asymptomatic bacteriuria. However, for patients with stones, obstruction, reflux, or other anatomical abnormalities, eradication of the organism is difficult. These patients are at risk of septicemia or recurrent urinary tract infections that are often caused by persistence of the same organisms.

Diagnosis—Microscopy of the urine for leukocytes suffices to diagnose a UTI in most circumstances. Ten or more leukocytes/high-powered field is considered abnormal. If complicated UTI is suspected or if infections have been resistant to therapy, culture and sensitivity of the urine are indicated. The culture is taken from a midstream, cleanvoided, urine specimen. The adequacy of the collection can be judged by the absence of squamous epithelial cells or multiple organisms on culture. Squamous epithelial cells and multiple organisms indicate contamination. If fewer than 1000 bacterial colonies/ml of urine are cultured, significant infection is not present as the bacteria are probably contaminants from the urethra or perineal areas. If between 10^3 and 10^4 organisms/ml of urine is cultured, interpretation depends on the apparent degree of contamination of the specimen and the plausibility of the organism cultured: often the culture should be repeated. If there are more than 10^5 organisms/ml of urine, the diagnosis of a UTI will be correct in 80% of the cases. When urine is obtained from the bladder, ureters, or renal pelvis by sterile technique, the presence of any number of bacteria indicates a UTI.

The extent of the diagnostic workup of a patient with a UTI depends on whether it is the first infection, the age and sex of the patient and the presence of underlying disease. Male children and adult men presenting with a second UTI should receive a complete evaluation to rule out anatomical abnormalities. A female in the childbearing years may be diagnosed on the basis of urinalysis and gram stain of bacteria in the urine if this is the first infection. Cultures are obtained in other situations. Recurrent UTIs require a complete diagnostic workup in certain circumstances.

Sexually Transmitted Disease

Sexually transmitted disease (STD) refers to a disease acquired through sexual activity. There are many diseases in this category, and STD refers to no specific one; therefore, the term is confusing.

GONORRHEA (GC)—This is an extremely common disease that is transmitted by genital, anal-genital, or oral-genital contact.

Epidemiology—Gonorrhea is pandemic in the US, particularly in poor urban settings. The highest incidence is in persons aged 15 to 24 years, minorities, and persons living in the Southeastern United States. Historically, gonorrhea has been reported more commonly in men. This difference has been ascribed to the higher prevalence of asymptomatic disease in women. However, since 1980 there has been a decreased incidence in homosexual men coupled with better case finding in women that has resulted in a near equal rate of disease for both men and women in the US currently.

Etiology—Gonorrhea is caused by the fastidious, nonspore-forming gram-negative diplococcus, *Neisseria gonorrhoeae*. This organism requires precise conditions for growth. It dies quickly on a dry swab, survives only briefly on a moist towel, and does not grow at room temperature. Certain strains of *N gonorrhoeae* are resistant to penicillin and tetracycline. There are no nonhuman reservoirs of gonorrhea.

Pathophysiology—After an infected person inoculates gonococci onto a mucous membrane, local invasion occurs. The hallmark of GC is copious, yellow pus. Common sites of inoculation are the pharynx, urethra, cervix, and anus. The incubation period for gonorrhea is 3 to 5 days. Once inoculated in the genital tract, the infection may ascend, particularly in the female. However, in men, epididymitis and prostatitis are rare. In the female, the gonococcus does not survive well in the uterus but does infect the fallopian tubes in about 15% of cases. This may cause scarring and later sterility. About 1–3% of affected adults develop gonococcemia; two-thirds of these are females. Distant sites of infection include joints, meninges, and heart valves.

Symptoms and Signs—Clinical manifestations of gonorrhea depend on the site and duration of infection and whether there has been local or systemic spread.

A profuse, purulent, yellow urethral discharge associated with dysuria and frequency develops in 90–95% of infected males. If untreated, the urethritis will resolve in 8 weeks. Anorectal infections are usually asymptomatic but may produce anorectal burning or itching, tenesmus, and a bloody, mucopurulent, rectal discharge. These symptoms may subside without treatment. Pharyngeal infections may produce an exudative tonsillitis but are most commonly asymptomatic.

Only 5–10% of infected females develop symptoms, which include dysuria, frequency, increased vaginal discharge, abnormal menstrual bleeding, and anorectal discomfort. The symptoms of urethritis may be confused with a urinary tract infection, and the increased vaginal discharge may be attributed to vaginitis. Vaginitis and UTI may occur concomitantly with GC. Lower abdominal tenderness and pain suggest pelvic inflammatory disease. Fever chills, nausea, vomiting, and leukocytosis may also occur. Physical examination reveals signs of pelvic peritonitis.

Gonococcemia may be the first sign of disseminated infection and includes fever, polyarthralgias and skin lesions that usually are located on the distal extremities. The skin lesions may be papular, petechial, pustular, hemorrhagic, or necrotic. Tenosynovitis or a septic arthritis of a single, large joint or of several joints usually follows but may precede symptoms and signs of gonococcemia. The synovial fluid is purulent, and joint destruction occurs very rapidly without proper treatment.

Diagnosis—Diagnosis of gonorrhea in a male is made on a gram stain of the urethral discharge by the presence of gram-negative diplococci within leukocytes. If gram-negative diplococci are seen but are extracellular, a culture is required for diagnosis. The diagnosis of gonorrhea in females is made by culture of the cervix. The anal canal and pharynx should also be cultured in women and homosexual men. Blood cultures are unlikely to be positive for gonococcus 48 hours after the onset of gonococcemia.

SYPHILIS—Syphilis is a chronic systemic infection that is seen in three stages: primary, secondary, and tertiary, that progress over many years. In untreated syphilis, degeneration eventually occurs in the central nervous and cardiovascular systems.

Epidemiology—The incidence of syphilis has increased over the past decade. The incidences of tertiary and congenital syphilis had been declining since 1943 but rose in epidemics in 1982 and 1990. As with other sexually transmitted diseases, syphilis is more common among indigent nonwhites living in urban areas, illicit drug users, homosexuals, and patients infected with HIV.

Etiology—Syphilis is caused by the spirochete *Treponema pallidum*, a spiral-shaped organism that is not seen under an ordinary light microscope but that can be visualized using the dark-field technique. The organisms have not been cultured because their growth requirements are so precise. The only naturally occurring host for *T pallidum* is man.

Pathophysiology—Nearly all cases of syphilis are acquired by sexual contact with infectious lesions. Syphilis may be acquired rarely by nonsexual personal contact, contact with contaminated fomites, blood transfusions, and in utero. The spirochete penetrates intact mucous membranes or abraded skin and enters the lymphatics and blood within a few hours. The average incubation time for syphilis is 21 days; however, it ranges from 10 to 90 days depending on the size of the inoculum.

The immune response to infection with *T pallidum* begins with the migration to the site of infection by polymorphonuclear cells that coincide with the formation and eventual resolution of the primary chancre. Antibodies also form that can be detected relatively early in most infected patients. Despite this immune response, without treatment widespread dissemination of organisms occurs which leads to the secondary and tertiary stages of the disease.

Symptoms and Signs—The hallmark of *primary syphilis* is the chancre. The chancre begins as a papule, which rapidly becomes eroded and forms an ulcer that is generally painless. The chancre is most commonly found on the external genitalia or the anal canal but can be located anywhere. This primary chancre heals spontaneously in 2 to 6 weeks. The chancre is highly infectious. Approximately 25% of patients with untreated disease will progress to secondary syphilis.

Secondary syphilis appears approximately 6 weeks after the chancre has healed and is characterized by appearance of nonpruritic red or pink macules on the trunk and proximal extremities. In about 1 to 2 months, red papular lesions also appear that may progress to pustular or necrotic lesions. The lesions are widespread and may involve the palms, soles, face, and scalp. The papules may scale, but vesicles are not

seen. Lymphadenopathy and headache are common. Just as in primary syphilis, the manifestations of secondary syphilis typically resolve spontaneously even in the absence of therapy. Occasionally, patients may experience relapsing secondary syphilis for up to five years after the initial episode.

Tertiary syphilis may occur 1 to 30 years after the primary infection. It is not necessary for individuals to have experience clinically symptomatic primary syphilis prior to the development of tertiary symptoms. The most common manifestations of tertiary syphilis are central nervous system involvement, cardiovascular disease, and gummatous syphilis.

Central nervous system involvement is manifested primarily in one of three ways and is seen in 5% of patients. First, meningovascular syphilis may erupt 5 to 10 years after the primary infection and involves inflammation of the pia and the arachnoid. There may be either focal or widespread symptoms. Second, general paresis reflects widespread parenchymal damage to the brain. It causes changes in personality, affect, intellect, judgment, orientation, calculating ability, and insight. There will likely be hyperactive reflexes, difficulty with speech, and small irregular pupils that react to near, but not to light. General paresis is seen about 20 years after infection. Third, tabes dorsalis is due to demyelination of the posterior columns, dorsal roots, and dorsal-root ganglia of the spinal cord. Symptoms and signs include ataxia, wide-based gait, foot-slap, paresthesias, and bladder disturbances. There may be impotence and loss of position, deep pain, and temperature sensations. Trophic degeneration of joints and ulcers of the feet may develop as a result of loss of pain sensation. Tabes dorsalis occurs 25 to 30 years after infection.

Cardiovascular syphilis generally begins 15 to 30 years after the initial infection. Classically the aorta is involved with resulting dilation of the aorta (aneurysm). The aortic valve may also become incompetent with regurgitation of blood through the weakened valve. The onset is generally insidious, and most patients present with an asymptomatic murmur or sometimes with congestive heart failure. The coronary arteries may also be compromised and coronary thrombosis may occur.

Gummas are granulomatous, nodular lesions that may occur anywhere but are most common on the skin or in the bones. Gummas involving the internal organs may appear as mass lesions. Gummas are rare but are being seen more frequently in HIV-infected individuals.

In HIV-coinfected people, persistent chancres, secondary infection, and early neurosyphilis may be more common.

GENITAL HERPES—Genital Herpes or herpes genitalis is a common sexually transmitted disease in the US. It occurs in acute (primary) and recurrent forms.

Epidemiology—Genital herpes has reached epidemic proportions in this country, and the rate of occurrence seems to be increasing. The peak incidence is during the sexually active years, although all age groups are affected. Herpes infections occur in all socioeconomic groups. Recurrent episodes may be more frequent than primary ones. Genital herpes is the most common ulcerative sexually transmitted disease.

Pathophysiology—Genital herpes is contracted primarily through sexual contact with an individual who has an active infection. The primary infection consists of grouped vesicles on an inflamed base. It is spread by lymphatics, blood, and ascending sensory nerves. The virus resides in dorsal root ganglia and periodically descends to the skin to cause lesions. Causes of reactivation from the latent stage are not clear. It is often difficult to identify primary cases of herpes genitalia, since many cases are asymptomatic. Recurrent episodes generally are shorter in duration, less severe, and less likely to be associated with systemic involvement than are primary cases.

Symptoms and Signs—A prodromal stage usually precedes the appearance of skin lesions. Symptoms during this phase may include pain, tingling sensations, or itching. Usually, within 24 hours, lesions appear that initially are papular and rapidly progress through vesicular, ulcer, and crusting stages in an otherwise asymptomatic patient. Systemic involvement may occur in neonates and immunocompromised patients.

A typical primary episode lasts 2 to 3 weeks, whereas recurrent cases are much shorter (5–10 days). Recurrent disease is more likely to occur in patients with a more severe initial episode, a prior recurrence, a history of sexually transmitted disease, younger age, and immunosuppression. While lesions often are limited to the genitals and perineal area, they also may occur on the thighs and buttocks.

Etiology—Herpes Simplex Virus Type 2 (HSV-2) causes the vast majority of cases of genital herpes. A very small number may be caused by Herpes Simplex Virus Type 1 (HSV-1) or a concurrent infection with both types. HSV is a DNA virus and is identified through cultures and serological testing for antibodies to the virus.

Diagnosis—The diagnosis of herpes genitalis is made through history, physical exam, and culture of scrapings or biopsies. Serological techniques can help in the diagnosis of a primary infection though these generally take 4 to 6 weeks to become positive.

Respiratory Tract Infections

These infections are the most common of acute illnesses. Etiologic agents include viruses, bacteria, *Mycoplasma*, and rarely other organisms. Lower respiratory tract infections usually indicate an impairment of host defenses.

Normal Anatomy and Physiology—A number of organisms normally colonize the nasopharynx (normal flora). Most of these organisms are not pathogenic and return after antibiotic therapy. Normal flora may inhibit growth of pathogenic organisms. Potential pathogens often colonize the upper respiratory tract; although they will not often result in infection to the individual, they may transmit disease to others (eg, meningococcus). Transient colonizers may become infectious in some individuals. Anaerobic organisms constitute 90% of the normal flora of the upper respiratory tract. Also, normal flora does not usually extend below the larynx. The lower respiratory tract is sterile in healthy people.

The lungs are protected from infection by several defense mechanisms. The lining of the respiratory tract is composed of sticky surfaces on which particles adhere. Particles larger than 5 μm are usually filtered efficiently and do not reach the alveoli. The lungs also have mechanisms to remove particles that reach the bronchi or alveoli. Coughing and sneezing are natural defenses for removing particles. Ciliated epithelial cells line the lower respiratory tract. Mucus secretion by goblet cells helps to trap particles and suspend them for transport by the cilia. This mucociliary transport system is the most important means for clearing particles. Macrophages located in alveoli can engulf particles. Specific antibodies and other soluble factors, such as lactoferrin and lysozyme, also contribute to clearance. If a particle cannot be removed or destroyed within the lung, a granuloma forms around it to wall it off.

Environmental factors such as air pollution, cigarette smoking, drugs such as alcohol and anesthetics, and various disease states such as congestive heart failure and leukemia can suppress the normal defensive mechanisms of the lung.

UPPER RESPIRATORY TRACT INFECTIONS—

Epidemiology—URIs follow seasonal variation, with the incidence being highest in winter and lowest in summer. This type of virus infection is transmitted mainly through the coughing and sneezing of infectious aerosolized droplets, but transmission occurs principally through contamination of hands and objects by nasal secretions and saliva. Infection depends on the size of the inoculum and the response of the host.

Etiology—Approximately 95% of upper respiratory tract infections are due to viruses. More than 150 serotypes, representing 12 groups of viruses, have been associated with URIs. Rhinoviruses cause 40% of respiratory illness; adenoviruses cause 2 to 10%, with the remainder being caused by respiratory syncytial virus (RSV), coronavirus, or influenza viruses. *Mycoplasma*, *Chlamydia*, and at times bacteria primarily cause the remaining 5% of URIs with Streptococci being the most common of these.

Pathophysiology—Respiratory viruses cause mucosal sloughing and consequently decreased lung defense mechanisms. This predisposes to serious bacterial infections, although this suprainfection occurs in only a minority of patients.

Symptoms and Signs—The symptoms and signs of a viral URI are familiar and are known as the "common cold." These include a coryzal syndrome characterized by nasal stuffiness and discharge, sneezing, moderate sore throat, and mild constitutional symptoms. Fever may or may not be present. Children with rhinoviruses may develop bronchitis, bronchiolitis, and pneumonia. Both children and adults with adenovirus, respiratory syncytial virus, or influenza viruses may develop lower respiratory infection.

Diagnosis—The diagnosis of URI is made on clinical grounds. Rarely is any additional testing required as these are primarily self-limiting viral infections for which there is no specific therapy. Occasionally, such as in cases of suspected influenza or RSV, nasal/throat washings for viral culture might be considered to confirm the diagnosis when antiviral therapy has been initiated.

STREPTOCOCCAL INFECTIONS—Streptococcal infections are important because of the seriousness of the acute illness as well as the late complications that are not infective but are mediated immunologically. Acute respiratory tract infections with streptococci may manifest as streptococcal pharyngitis, scarlet fever, or pneumonia. The late complications include acute rheumatic fever, rheumatic heart disease, and acute glomerulonephritis.

Epidemiology—Streptococcal infections occur throughout the population. Respiratory streptococcal infections are more common during the colder months. Scarlet fever is usually a disease of children between 6 months and 10 years. Infants less than 3 months rarely have streptococcal infections. Streptococcal pharyngitis occurs commonly among children and young adults. As many as 20% of the population are asymptomatic carriers of Group A streptococcus. A streptococcal URI may be spread by inhalation of respiratory secretions. Epidemics of streptococcal URIs occur.

Etiology—Streptococci are gram-positive cocci that tend to form chains. Three groups have been identified by their ability to hemolyze red cells in culture media by the enzymes streptolysin O and S. Alpha streptococci (or viridans streptococci), beta hemolytic streptococci, and gamma-nonhemolytic streptococci are the three groups. There are 13 serologic types of streptococci designated by the letters A to O. Most bacterial URIs are caused by Group A streptococci. The late complications of rheumatic fever and glomerulonephritis have been attributed only to Group A streptococci.

Pathophysiology—Streptococci are inhaled into the nasopharynx and normally are cleared by defense mechanisms or become transient colonizers. The size of the inoculum, the virulence of the organism, the presence of type-specific immunity, and the defense mechanisms of the host determine if an infection is to occur. Type-specific immunity lasts for years.

Symptoms and Signs—Streptococcal infections present a variable clinical syndrome, and as many as 40% of individuals may be asymptomatic. The incubation period usually lasts 3 to 5 days. The onset is acute, and the illness includes fever, chills, headache, sore throat, anorexia, malaise and, in children, nausea and vomiting. Symptoms reach a maximum in 1 to 2 days. Swallowing worsens the sore throat, hoarseness is present, and nasal stuffiness, nasal discharge, and a nonproductive cough may occur. Earache is common. Scarlet fever is streptococcal pharyngitis followed by a rash with circumoral pallor.

Patients with streptococcal pharyngitis may be mildly to moderately ill, with fever to 40°. Tachycardia and a diffusely red posterior pharynx and soft palate are common. The uvula is edematous. Characteristically, there is an exudate on the tonsils, which may be scraped off without bleeding. The nasal discharge is thick, mucopurulent, and may contain blood.

The clinical course of a streptococcal URI is short with the fever resolving in 3 to 4 days or 5 to 9 days in adults and children, respectively. If scarlet fever develops, exfoliation of the epithelium begins as the rash fades.

Diagnosis—A positive rapid optical immunoassay or throat culture for Group A beta-hemolytic streptococci in the setting of the characteristic history, symptoms and signs makes the diagnosis of streptococcal pharyngitis.

PNEUMONIA—Pneumonia is an infection in the alveoli that only occurs when impairment of host defenses allows the organism access to alveoli and the infectious process cannot be contained. Pneumonia occurs more frequently in individuals with underlying chronic cardiopulmonary disease, immunologic compromise, habitual cigarette smoking, or alcoholism, although it is not uncommon is otherwise healthy individuals. It is also more likely in individuals who recently had a viral pneumonia or general anesthesia.

Epidemiology—Community-acquired pneumonia is the 6th most common cause of death in the US and the most common infectious cause with approximately 6 million cases annually. Approximately 5–60% of the population are asymptomatic carriers of the pneumococcus, depending on the season. The infection is more prevalent in winter and spring. Nearly 500,000 patients with pneumonia are admitted to hospital each year with more than half of these being patients over the age of 65 years. However, the incidence of pneumonia is actually highest among persons younger than age 65 years. The incidence of pneumococcal pneumonia has changed little, although the mortality has decreased greatly with the advent of antibiotics except in the elderly where the mortality rate continues to rise. *Pneumococcus* accounts for 30–60% of community-acquired for which a cause is found. Atypical organisms are felt to be present in up to 25% of cases and may be found as single causative organisms but not infrequently can also be found as mixed infections with other bacteria. *Staphylococcus* is uncommon as a cause of pneumonia except in patients who are hospitalized or those who have had influenza recently. Anaerobic bacteria are often the causes of pneumonia in patients with impairment of swallowing who aspirate oral contents into the lung.

All causes of pneumonia are more frequent in patients with underlying lung disease such as chronic bronchitis or emphysema.

Etiology—Pneumonia may be caused by bacteria, atypical organisms (*Mycoplasma*, *Chlamydia*), or viruses. Rarely fungi are causes of pneumonia. Bacterial causes of pneumonia are the most common. *Pneumococcus* (*Streptococcus pneumoniae*), the most frequent bacterial cause of pneumonia, is a gram-positive encapsulated coccus that usually grows in pairs, hence, the name diplococcus. Other common bacterial causes are *Haemophilus influenzae, Moraxella catarrhalis,* enteric gram-negative organisms, *Staphylococcus aureus, Legionella pneumoniae,* and anaerobic bacteria. However, a specific etiology can be identified in approximately 50–60% of cases.

Pneumonia is broadly divided into community-acquired pneumonia and hospital-acquired pneumonia. This division allows for differentiation in terms of expected pathogens, diagnostic approach, expected mortality, and treatment.

Community-acquired pneumonia has traditionally been divided into broad categories of "typical" verses "atypical" pneumonia. This classification was based on clinical presentation, patient demographics, and chest x-ray findings and was felt to provided clues to likely underlying pathogens. However, recent data has shown that this classification is inaccurate and not useful.

Pathophysiology—Bacterial pneumonia occurs when pathogens are aspirated or inhaled into the lungs and the normal defense mechanisms fail in their ability to promptly remove the offending organisms. Bacteria that are aspirated into the lung and usually lodge in the right-lower, right-middle, or left-lower lobe, where they multiply rapidly. The response to the multiplying organisms involves transudation of fluid into the alveoli, which becomes a growth medium for the organism and a mode for local spread to other alveoli, segments, lobules, lobes, and pleura. Polymorphonuclear leukocytes migrate to the site of infection to phagocytose the bacteria. Macrophages appear later to clean up the fibrin and debris. Antibodies against the *Pneumococcus* or other bacteria enhance phagocytosis and cause organisms to agglutinate and adhere to the alveolar wall, thus slowing spread of the infection. Bacteremia is usually transient. The most common complication is the migration of infection to the pleural space causing the formation of an empyema. Lung abscess formation and spread of infection to distant sites such as meninges, pericardium, or joints are other complications but are less common.

Symptoms and Signs—The clinical course of pneumococcal pneumonia is classic. A URI syndrome may precede the pneumonia by a few days. The onset is abrupt, and patients often can state the hour of onset. In 80% of patients, there is a sudden shaking chill and a rapid rise in temperature with tachycardia and tachypnea. In 75% of patients, pleuritic chest pain and a productive cough develop. The sputum is mucoid and pink or rusty in color. Dyspnea is a common complaint. The patient will appear acutely ill but will not complain of nausea, headache, or malaise. If untreated, the symptoms and signs last for 7 to 10 days. Then there is diaphoresis, a sudden drop in temperature, and dramatic improvement. Circulatory collapse and heart failure are common in fatal cases. With other bacterial causes of pneumonia, the onset of symptoms may be more insidious, but fever, productive cough, and dyspnea are still typically present regardless of the underlying pathogen.

On physical examination, breath sounds are decreased, and crackles and rhonchi are present. The chest radiograph shows a homogeneous density in the affected areas. There is a leukocytosis with 70–90% of the WBC being mature or immature polymorphonuclear leukocytes, the "shift to the left." Blood culture is positive in only 10–20% of cases. Gram stain of the sputum shows many PMNs and gram-positive cocci usually in pairs in cases of pneumococcal pneumonia. The Gram stain is less sensitive and specific for other bacterial etiologies.

Poor prognostic signs include leukopenia, bacteremia, multilobar involvement, extrapulmonary infection, underlying systemic disease, and circulatory collapse. The fatality rate in pneumococcal pneumonia is about 5% despite appropriate treatment.

Diagnosis—The diagnosis of pneumonia is suspected by the clinical findings of cough, fever, and dyspnea. A chest x-ray confirms the clinical suspicion. No other specific tests are available for diagnosis. However, an elevated WBC, sputum Gram stain, blood culture, serological tests, and low oxygen saturation can assist in determining prognosis and/or etiology.

MYCOPLASMA—Mycoplasma pneumoniae, (previously called pleuropneumonia-like organisms (PPLO) or Eaton's agent), causes asymptomatic infections, upper respiratory infections, and pneumonia. Mycoplasma pneumonia has been called atypical pneumonia or walking pneumonia, to distinguish it from pneumococcal pneumonia, but the clinical distinction is not crisp.

Epidemiology—The infection is spread by inhalation of respiratory secretions and is characterized by occurrence among many family mem-

bers or in large numbers of people living in crowded environments such as military bases and college dormitories. *Mycoplasma* infections are common among children and young adults. Traditionally *Mycoplasma* has been felt to be rare in older adults. However, recent data shows that the incidence rises consistently with age. *Mycoplasma* pneumoniae accounts for 15–20% of all pneumonias. *Mycoplasma* infections are more common in the winter.

Etiology—*Mycoplasma* is a unique organism of extremely small size. Instead of a cell wall, a unit membrane surrounds each *Mycoplasma*. Lacking cell walls, *Mycoplasma* resistant to β-lactam antibiotics. *Mycoplasma* frequently may be found as normal flora in the upper respiratory tract.

Symptoms and Signs—The incubation period for *Mycoplasma* varies from 9 to 12 days. The disease begins as a URI that progresses to bronchitis and subsequently to pneumonia in 3–10% of cases. A nonproductive cough is the most characteristic symptom. In cases of pneumonia, the cough may become productive of blood-tinged sputum later in the course. Headache, general malaise, muscle aches, nasal congestion, and sore throat are common.

The clinical course of the disease is variable. Fever may persist for 2 weeks in untreated cases. The pneumonia is usually multilobar and may be bilateral. Lower lobes are involved more commonly than upper lobes. The infiltrate is less dense than in bacterial pneumonia and often is of an interstitial pattern. The physical findings on chest examination usually are much less striking than the severity of disease noted on the chest radiograph.

Complications are rare even without treatment.

Diagnosis—The diagnosis, as with other forms of pneumonia, is based on the history and clinical picture and the chest radiograph. There is a minimal increase in WBC count without a "shift to the left." Lymphocytosis with atypical forms may be present. Cold agglutinins are positive in 50% of the cases after the second week of the illness. A rise in specific antibodies to *Mycoplasma* is a more sensitive and specific test. It takes 2 to 4 weeks to culture *Mycoplasma*.

SEVERE ACUTE RESPIRATORY SYNDROME (SARS)—

Epidemiology—In early 2003, a new, virulent, and apparently highly contagious URI appeared in Guangdong Province of China and has spread to other countries in South East Asia and to Toronto, Canada.

Etiology—The causative agent appears to be a coronavirus.

Pathology—Diffuse alveolar damage and consolidation.

Symptoms and Signs—Patients present with high fever, dry cough, rigor, dyspnea, malaise, and headache. Examination of the chest shows crackles and dullness to percussion. Lymphopenia is common. The chest x-ray shows progressive consolidation.

Diagnosis—Diagnosis is based on the clinical picture in a patient who has been in an endemic area or who has had contact with known cases. Serological tests and culture methods are being developed.

Tuberculosis

Tuberculosis (TB) is a bacterial infection that has greatly decreased in prevalence in the US but remains a threat. Although TB can involve many organs, pulmonary TB is the most common.

Epidemiology—Since the beginning of the 20th century the incidence of TB has been declining in the US, but this decline was punctuated by an increase around 1980 when there was an influx of refugees from Indochina. A second rise of about 20% followed in 1985–92, largely in HIV-infected people. In 1997, under 20,000 cases were reported, an all-time low. Microepidemics often occur in nursing homes and families.

Tubercle bacilli are aerosolized as droplets during coughing by a person with cavitary disease. After evaporation, droplet nuclei, which are 1 to 5 μm in diameter, can reach the alveoli and establish an infection in a susceptible host. The infectivity of a patient is related to the severity of the disease, the number of bacilli in the lesion, and the closeness and length of the contact. An infected person is considered no longer contagious after about 2 weeks of appropriate chemotherapy.

Etiology—*Mycobacterium tuberculosis* is a rod-shaped organism that requires high oxygen tension for optimum growth and produces no toxins or enzymes. The organism has unique staining properties due to the lipid composition of the cell wall. Carbol-fuchsin stain does not wash off with acid, hence the name "acid fast." The bacilli can be cultured.

Pathophysiology—Tubercle bacilli are inhaled and deposited in peripheral alveoli throughout the lung. Before the infection can be contained by a local cellular response, the bacilli are drained to lymph nodes in the hilum and then disseminated throughout the body by the bloodstream.

Sites that are seeded by bacilli include the apices of the lungs, the kidneys, the growing ends of bones, and other areas of high oxygen tension. Cellular immunity involving lymphocytes, macrophages, and giant cells develops in several weeks. Once cellular immunity develops, the reaction forms granulomas at the sites of infection, and in time caseous necrosis may develop in these granulomas. During caseation, cytotoxic material released from T lymphocytes destroys the bacilli as well as the surrounding tissue. The sites then heal by resolution, fibrosis, and calcification. In some cases the immunity is inadequate, and overwhelming infection develops. Healed lesions still contain viable tubercle bacilli. These may remain dormant for the life of the individual. In 10% of cases, these lesions develop into clinical disease sometime after the initial infection.

Symptoms and Signs—The initial infection of primary TB usually produces few symptoms or signs. The incubation period is 4 to 8 weeks. Mild fever and malaise may occur as tuberculin hypersensitivity develops. In some cases, especially in a child less than 3 years, an overwhelming infection may result from the primary infection.

Pulmonary tuberculosis usually occurs after a period of dormancy in a previously infected individual. The onset is insidious. The patient may be asymptomatic with a routine chest radiograph leading to the diagnosis. Fever to 40° may occur in the late afternoon or evening. Night sweats are common. General malaise, fatigue, irritability, and weight loss may occur. A cough, productive of green or yellow sputum that may be blood-streaked is common. When cavitation occurs, highly infectious material spills into the bronchi and is coughed up.

Spread of pulmonary tuberculosis to the pleura results in pleuritic chest pain and the formation of a pleural effusion as part of the inflammatory reaction. The presence of a large effusion may compromise lung function and result in the complaint of dyspnea. Tuberculosis also can spread from the lung or the lymph nodes into the pericardium where the same inflammatory process occurs. A friction rub may be heard. Later, the inflamed pericardium may scar down, calcify, restrict cardiac motion, and present as congestive heart failure.

During the dissemination phase, bacilli are seeded in the kidneys, bone, adrenals, and meninges. At each site the same inflammatory process occurs with caseation and liquefaction. If the infection cannot be contained, local spread may occur. Tuberculosis in the kidneys may result in infection of the rest of the genitourinary tract and present as cystitis, epididymitis, or prostatitis. In females, tuberculosis of the fallopian tubes and uterus may result in abdominal pain, vaginal discharge, sterility, or ectopic pregnancy. Spondylitis may result in localized back pain or compression of the spinal cord. Tuberculosis of the adrenal glands may cause total destruction of the glands and result in Addison's disease. Tuberculous meningitis also is seen. Symptoms and signs include headache, restlessness, irritability, nausea, vomiting, and stiffness of the neck. A change in mentation may be the only sign of the disease.

Miliary tuberculosis is a massive dissemination of tubercle bacilli throughout the body. Lesions are found in the liver, spleen, bone marrow, and other organs (which do not have a high oxygen tension) in addition to the previously mentioned sites of typical spread. The symptoms and signs are nonspecific and include dyspnea, weight loss, weakness, fever, night sweats, and gastrointestinal disturbances. Death is certain unless appropriate treatment is instituted promptly.

Diagnosis—The diagnosis of tuberculosis rests on the use of a skin test with tuberculin, which is the protein fraction of the tubercle bacilli. However, this test cannot discriminate between dormant and active disease. Sensitized lymphocytes accumulate at the site of intradermal injection of tuberculin. Five tuberculin units are injected and the skin test is read in 48 to 72 hours. The criterion for a positive test depends on the age of the patient, degree of exposure, and HIV status. False negative tests occur in 15–20% of patients with clinical tuberculosis. The skin test does not become positive until the development of cellular immunity. In patients with a decreased number of lymphocytes, an overwhelming infection, a pleural effusion, or a fever, the skin test may be falsely negative. The chest radiograph also is essential to the diagnosis of pulmonary tuberculosis. It shows multinodular infiltrates, with or without cavitation, in one or both upper lobes. The Ziehl-Neelsen stain for acid-fast bacilli has been largely supplanted by fluorescent staining methods and nucleic acid amplification techniques: the latter require less than 6 hours, but are expensive. Sputum also may be cultured for the organisms. With modern BACTEC radiometric culture systems, growth can often be detected within 10 days.

Nontuberculous Mycobacterial Disease

Many mycobacteria live freely in our environment and are not generally pathogenic unless host defenses are impaired. They have become important causes of disease in patients with AIDS.

Epidemiology—*M avium* is ubiquitous and is particularly found in water sources and wet environments. *M kansasii* is concentrated in the urban midwest of the US. Person-to-person transmission has never been shown, but infection is extremely common. Skin testing indicates that at least 40 million Americans have been infected, but few of these became ill.

Etiology—*Mycobacterium avium* differs from *M tuberculosis* in growth rate, colony morphology and pigment formation, DNA composition, and pathogenicity. It is readily seen with conventional acid-fast staining.

Pathophysiology—Pulmonary infection is the most common site. Infection is presumably by inhalation. Generally the disease progresses slowly, but occasionally it advances rapidly. In patients with AIDS dissemination is common.

Symptoms and Signs—Patients with AIDS usually present with fever. The liver and spleen may be enlarged. The organism can be readily grown from many sites including blood. Organisms grow within 5 days in appropriate liquid media, and DNA probes permit rapid species identification.

Infections Of the GI Tract

VIRAL HEPATITIS—See Gastroenterology section.

INFECTIOUS DIARRHEA—

Normal Anatomy and Physiology—The gastrointestinal tract has several defenses against infection. Gastric acid keeps the stomach sterile. If intragastric pH increases, fewer pathogens are needed to establish an infection. The remainder of the gastrointestinal tract has a normal bacterial flora that inhibits the growth of other organisms. The flora of the large intestine is composed predominantly of anaerobes. Some species of the normal flora produce short-chain fatty acids or antibiotics such as clostin that prevent the growth of pathogens. Other members of the normal flora compete with pathogens for nutrients. Antibiotics that suppress normal flora predispose to bacterial infection. Cells that produce mucus line the gastrointestinal tract. This mucus forms a barrier to bacterial invasion of the gut wall. Locally produced IgA antibodies and antibodies produced elsewhere, such as IgG, enhance phagocytosis of bacteria within the GI tract. Motility of the gastrointestinal tract moves organisms out and thus prevents infections. Diarrhea increases transit and rids the body of organisms. Antimotility agents interfere with this defense.

Diarrhea is defined as an increase in numbers of stools per day and/or an increase in stool volume. Acute diarrhea is sudden in onset, lasts for less than 2 weeks, and usually is caused by an infectious agent. Chronic diarrhea is of longer duration and usually is due to noninfectious gastrointestinal disease.

Epidemiology—The transmission of the causative agent is by the fecal-oral route in most cases. Contaminated objects, hands, food, and water may transmit the agent. The incidence of infectious diarrhea in the general population has been estimated to be approximately 20 to 40 cases per 100 person years. Foodborne diseases account for roughly 76 million cases, 325,000 hospitalizations, and 5000 deaths in the US annually. These rates are most likely underestimates, as many patients do not seek medical attention.

Etiology—Bacterial toxins, bacterial organisms, viruses, or parasites may cause diarrhea. Diarrhea need not be caused by a pathogen but may be due to changes in normal flora or by normal colonic flora reaching the small intestine. Bacteria that commonly cause diarrhea by the production of toxins include enterotoxigenic *Escherichia coli*, *Staphylococcus*, *Clostridium perfringens*, and *Clostridium difficile*. Bacterial diarrhea is caused by *Shigella, Salmonella, Campylobacter jejuni,* and *Yersinia enterocolitica* in the US and *Vibrio cholerae* in other countries. Reovirus-like agent (Norwalk agent), echo, and coxsackie viruses commonly cause diarrhea, whereas influenza viruses do not. Parasites include *Entamoeba histolytica* and *Giardia lamblia* as common causes of diarrhea. The frequency of identifying an organism is 2–40%.

Pathophysiology—Viral diarrhea may cause villous shortening in the small intestine, an increase in the number of crypt cells, and widening of the lamina propria. Diarrhea caused by bacterial invasion results in hyperemia, leukocyte infiltration, and frank ulceration of the bowel wall. *Entamoeba histolytica* produces an inflammatory colitis similar to ulcerative colitis except for the presence of the parasite and larger, flask-shaped ulcers of the colonic mucosa. Bacteria may cause diarrhea via enterotoxin-induced hypersecretion or invasion of the gut wall by the bacteria. Enterotoxins stimulate adenyl cyclase in the mucosal cells of the intestine that results in massive secretion of fluid and electrolytes into the bowel lumen. Mucosal integrity is preserved and absorption is normal. In bacterial invasion, the damage to the mucosa results in defective absorption. Giardia probably produces diarrhea by the same mechanism since invasion of the small bowel occurs. *C difficile* causes pseudomembranous colitis (antibiotic-associated colitis).

Symptoms and Signs—Systemic symptoms including fever, headache, anorexia, vomiting, malaise, and myalgias may accompany diarrhea regardless of the etiology except when toxins are ingested.

Twelve to 24 hours after eating food contaminated by *Clostridium perfringens* or *Staphylococcus*, diarrhea with abdominal pain, cramps and nausea, but no vomiting or systemic symptoms occurs. The diarrhea contains no pus or blood. Recovery occurs in 12 to 24 hours.

In diarrhea in which the mucosa is invaded by organisms, such as Shigella or Salmonella, systemic symptoms occur along with lower abdominal cramps, tenesmus, and rectal urgency. Pus and erythrocytes or gross blood are found in the stool. Shigella causes explosive diarrhea and fever. The disease is usually self-limited with the fever subsiding in 4 days and the diarrhea subsiding in 1 wk. Shigella also produces a neurotoxin that may cause seizures in children. Salmonella produces a less acute clinical picture.

Enterotoxigenic *E coli*, frequently the causative agent in "turista" or "traveler's diarrhea," may produce mild or severe symptoms. The incubation period is 24 to 48 hours, and the diarrhea lasts for 2 to 7 days. The stools contain no blood and few white blood cells.

Nausea and vomiting and other systemic symptoms usually accompany viral diarrhea. The diarrhea is usually mild, recovery occurs in 48 hours, but malabsorption due to lactase deficiency may persist for several weeks. No red blood cells or white blood cells are seen in examination of the stool.

The prognosis of acute infectious diarrhea is usually excellent when treated with adequate fluid replacement. Complications are rare except in infants or extremely debilitated patients who are unable to tolerate the dehydration. Pseudomembranous colitis usually responds promptly to discontinuation of the causative antibiotic, although some cases require treatment with an antibiotic directed at *C difficile*.

Diagnosis—Diagnosis of the specific cause of infectious diarrhea is frequently made on clinical grounds alone as the majority of cases are self-limited. In cases where there is persistent fever, bloody diarrhea, or symptoms lasting for more than 4 days, a stool culture can be useful. Stool examination for the presence of WBCs is also beneficial to exclude non-infectious or non-inflammatory causes. With the proper history, stool analysis for *C difficile* toxin is beneficial as there is specific therapy for this condition. Likewise, when suspected by history, a stool examination for ova and parasites may yield a specific diagnosis.

Central Nervous System Infection

Meningitis and encephalitis are medical emergencies requiring rapid diagnosis and specific therapy. While meningitis involves only the leptomeninges, encephalitis involves the brain tissue itself and also may involve the meninges.

Normal Anatomy and Physiology—The central nervous system is composed of the brain and spinal cord. These structures are enveloped by the meninges, a three-layered fibrous structure within which flows the cerebral spinal fluid. This is a closed structure and is normally sterile.

Epidemiology—There are between 30,000 and 40,000 cases of meningitis annually in the US. People of any age may become infected; however, the frequency of infection and the type of organism varies with age. The highest incidence is found in neonates who are primarily infected with Group B streptococcus during the birth process. Gram-negative bacteria, enterococci, and *Listeria monocytogenes* are also seen. From age 1 month to 23 months *Streptococcus pneumoniae* and *Neisseria meningitides* are the most common organisms. From age 2 to 8, *N meningitides* accounts for more than half of all cases with *S pneumoniae* being second in frequency. *Haemophilus influenzae* type b used to have a high rate of infection in this age group but vaccination has dramatically curtailed this organism. In adults up to age 60, *S pneumoniae* and *N meningitides* are most common. Over age 60, *S pneumoniae* still accounts for most cases but *Listeria monocytogenes* is also common.

Meningitis is most common in the winter and spring. The lowest incidence is noted in the summer months. Epidemics are uncommon in the US but are still seen in developing countries worldwide.

Etiology—Bacteria often cause meningitis with the most common pathogens in most age groups being *Streptococcus pneumoniae, Haemophilus influenzae*, and *Neisseria meningitidis*. In neonates, *Escherichia coli* and Group B streptococci are common. Viruses such as enteroviruses and mumps virus also cause meningitis. Fungal meningitis occurs predominantly in immunocompromised patients. Viruses such as mumps and herpes viruses are the usual cause of concomitant encephalitis.

Pathophysiology—Meningitis most commonly results from the colonization of the nasopharynx with potential pathogens, which gain

access to the CNS by mucosal invasion. Other causes are direct extension of bacteria through skull fractures caused by trauma. Also, hematological spread of bacteria associated with cases of endocarditis or UTI for example is known to occur. Immunocompromise from HIV, asplenia, corticosteroid use, etc. predisposes to meningitis.

Bacteria that cause meningitis have the ability to invade the mucosa of the nasopharynx as opposed to other normal flora. This leads to transient bacteremia. Generally, the immune system is able to clear these bacteremic episodes before infection begins. However, circumstances that allow for large numbers of bacteria to invade and escape rapid clearance can lead to infection of the CSF. When infection of the CSF occurs, the organisms multiply rapidly. This initiates an intense inflammatory host response. It is the inflammatory response that is responsible for most of the symptoms and signs of meningitis that are seen clinically.

Symptoms and Signs—Systemic manifestations of CNS infection include fever, irritability, and somnolence. A single seizure prior to diagnosis is not uncommon. Nuchal rigidity and headache are often present. The headache is usually described as severe and generalized. Frequently there is hypersensitivity to light and/or sounds. Nuchal rigidity may not be a specific complaint but usually can be found on physical exam. It is manifested by inability of the patient to touch their chin to their chest with either passive or active flexion of the neck. Changes in mental status, seizures, and other focal neurological signs can be seen but usually are later findings.

Diagnosis—The definitive diagnosis of meningitis usually depends on analysis of CSF obtained through lumbar puncture. Every patient should have a lumbar puncture unless contraindicated. Of note however, antibiotic therapy should be given promptly in suspected cases even if the lumbar puncture has not been obtained. In classical bacterial meningitis, CSF glucose is decreased, protein is increased, and white blood cells (predominantly polymorphonuclear cells) and bacteria are present. These findings are quite variable in viral and fungal meningitis, however. Blood cultures are positive in at least half of the cases of bacterial meningitis and can be useful diagnostically especially when CSF cannot be obtained prior to the administration of antibiotics. Similar CSF findings also may be present with encephalitis. These patients often have more severe CNS dysfunction with symptoms such as coma and paresis. A culture of brain material obtained through biopsy is often necessary to identify clearly the etiologic agent in encephalitis.

Infective Endocarditis

This is an infection of the heart valves or the endocardial lining of the heart wall. The etiologic agent is most commonly bacterial but may be fungal. Based on the clinical course the endocarditis is said to be either acute or subacute (duration greater than 6 weeks).

Normal Anatomy and Physiology—The heart valves are fibrous tissue structures that have no intrinsic blood supply. As a result, infection of these valves does not generate a host immune response to the infection (ie, migration of PMNs to the site of infection). As a result, antibiotics are essentially the only treatment for endocarditis. Bacterial or fungal infection of the valves form clusters of organisms known as vegetations that can often be seen by echocardiography.

Epidemiology—A total of 10,000 to 15,000 new cases of infective endocarditis occur annually in the US. There is a male predominance for endocarditis. More than half of patients are in over the age of 60. Endocarditis is uncommon in children. Risk factors for the development of endocarditis are IV drug use, prosthetic heart valves, and structural heart disease (especially rheumatic heart disease).

Etiology—Endocarditis can be classified in three categories: native valve endocarditis, endocarditis in IV drug abusers, and prosthetic valve endocarditis. These categories are associated with different infecting organisms. Endocarditis can also be classified as acute or subacute. Acute disease is caused by *Staphylococcus aureus* infecting native, normal heart valves. It is aggressive and rapidly destructive. It is fatal within 6 weeks if not treated. Subacute endocarditis is usually caused by viridans streptococci on damaged heart valves and is much more indolent in its course.

Native valve endocarditis may be caused by any organism but most commonly is due to viridans streptococci (55%), enterococci, and *S aureus*. Most patients will have some prior damage to the heart valves (eg, rheumatic, age-related degeneration); however, *S aureus* can attack normal valves.

Endocarditis in IV drug users is due to *S aureus* in more than 50% of the cases. Streptococci, enterococci, gram-negative organisms, and fungi also are seen. Polymicrobial infections can also be seen in this population. Unlike native valve endocarditis, IV drug users have infection of the tricuspid valve 50% of the time.

Prosthetic valve endocarditis is divided into early-onset (<60days from placement of the prosthetic valve) and late-onset (>60 days). *S epidermidis* and *S aureus* constitute more than 50% of the cases of early-onset disease with gram negatives and fungi also being common. Late-onset disease is most frequently caused by streptococci or other organisms that are indigenous flora.

A group of gram-negative fastidious organisms known as the HACEK group (Haemophilus, Actinobacillus, Cardiobacterium, Eikenella, Kingella) are responsible for "culture negative" endocarditis. Although it is possible to grow these organisms in culture, special growth conditions are generally required making the diagnosis more difficult.

Pathophysiology—In subacute endocarditis, the congenital or acquired abnormal valve causes flow disturbances that injure the endocardial lining of the heart valves or wall. This area of injury becomes a focus of thrombus formation, which is seeded with bacteria during transient periods of bacteremia. Dental work or manipulation of either the gastrointestinal or genitourinary tracts with endoscopes, catheters, and surgical instruments can cause bacteremia with indigenous flora. The mass of adherent thrombus and organisms is known as a vegetation. Vegetations grow, erode the valve, and may create myocardial abscesses. Fragments can break off as emboli. Acute endocarditis results from direct attack of normal valves by aggressive organisms that can destroy valves rapidly. Abscess formation and disruption of cardiac conducting tissue is more common in acute endocarditis. Emboli are also more common in acute endocarditis especially with *S aureus*.

Endocarditis is associated with injury to many organs. The pathophysiology involves emboli (both septic and sterile) from the heart focus and immune complexes. In the setting of chronic infection with continued stimulation of the immune system, immune complexes of antibody and antigen form and deposit in various organs, thereby initiating a potentially harmful inflammatory response. Some manifestations of emboli and immune complex deposition are described in the following section.

Symptoms and Signs—Subacute endocarditis often begins with non-specific constitutional complaints. Fever, sweats, anorexia, malaise, myalgias, and arthralgias are prominent. These symptoms often persist and the patient may receive several courses of antibiotics, a practice that interferes with correct diagnosis.

A previous heart murmur may change or a new murmur may occur. Petechiae may appear in the optic fundi, conjunctiva, mucosal surfaces, or skin. Subungual splinter hemorrhages are a feature of this disease as are peculiar lesions on the hands and fingertips (Janeway lesions and Osler's nodes). In acute endocarditis, skin pustules occur. Arthritis and osteomyelitis, splenomegaly, and retinal lesions may develop. Renal manifestations (flank pain, hematuria) may be secondary to renal infarction by emboli or immune complex-mediated glomerulonephritis. Pulmonary infiltrates caused by septic emboli may occur with right-sided endocarditis. Cardiac conduction defects or congestive heart failure may develop as the infection erodes into the conduction system or chordae tendineae, respectively. Stroke, seizures, or meningitis resulting from emboli are seen more commonly in patients with acute bacterial endocarditis.

Diagnosis—No single test makes the diagnosis of endocarditis. Rather, a constellation of findings from history, physical exam, blood cultures, and echocardiogram are required. Of these however, positive blood cultures are of paramount importance. Culture negative endocarditis is possible but is rare.

Because of the difficulty in making an accurate diagnosis, criteria have been established to assist clinicians with suspected cases. These criteria are divided into major and minor criteria. The presence of two major or one major and three minor or five minor criteria is highly associated with endocarditis. The major criteria focus on the presence of multiple positive blood cultures and echocardiogram evidence of valvular vegetations or paravalvular abscess.

Acquired Immunodeficiency Syndrome

This syndrome (AIDS) is a condition characterized by the development of life-threatening opportunistic infection or malignancies in a patient with severe depression of the T-cell-mediated immune system caused by infection with human immunodeficiency virus (HIV). AIDS was first described as a specific entity in the US in 1981, and its frequency and mortality since have increased geometrically. As of December 2000, a total of 58 million individuals have become infected with HIV,

and 21.8 million have died since the beginning of the epidemic. A total of 36.1 million are now living with HIV/AIDS, and 90% of these persons are living in developing countries with over 25 million in sub-Saharan Africa alone. Half of those infected are women and less than 25 years of age. In the year 2000, 5.3 million people became newly infected (2.2 million women and 600,000 children). It is estimated that worldwide there are 15,000 people who become infected daily.

Epidemiology—In the US, AIDS was first described in 1981 in previously healthy homosexual men with *Pneumocystis carinii* pneumonia and Kaposi's sarcoma. In the US, the number of cases of AIDS has risen steadily over the past 20 years to over 750,000 total cases and 430,000 deaths. While the incidence of new cases peaked in the late 1980s, there are still approximately 45,000 new cases annually. The number of persons currently living with HIV infection or AIDS in the US is estimated to be approximately 920,000 and 320,000, respectively. Historically most cases have been in homosexual men and intravenous drug users, with different proportions in different areas. Currently, the proportion of cases is still most common in homosexual men. However, the number of new cases infected heterosexually has now surpassed the number infected by injection drug use.

Advances in the treatment of HIV have caused a marked reduction in the number of deaths in the US and Western Europe. From 1996 to 1999 the number of deaths due to AIDS decreased by 50%. However, this trend slowed from the latter part of 1998 through 2000.

Etiology—AIDS is a syndrome that results from infection with either the HIV-1 or HIV-2 virus. HIV-1 was discovered to be the causative agent for AIDS in 1984, 3 years after the first reports of the disease. In 1986, a second type of HIV called HIV-2 was isolated from AIDS patients in West Africa. Both HIV-1 and HIV-2 have the same mode of transmission and cause the same immunodeficient syndrome. However, persons infected with HIV-2 seem to develop immunodeficiency more slowly and have a milder clinical syndrome. There are only a few reported cases of HIV-2 in the US at this time.

The human retroviruses all share certain important functional features. Like all retroviruses, they produce reverse transcriptase that produces a DNA copy of the RNA material of the virus. They also are unusually trophic for T4 lymphocytes. All retroviruses that cause human disease tend to live silently within their target cells until they are activated to replicate. The HIV viruses, but not the other human retroviruses, attach to the CD4 receptor of their target cells, which are principally T4 lymphocytes and monocytes/macrophages.

Pathophysiology—HIV has been isolated from the blood, semen, vaginal fluid, urine, and tears of AIDS patients. Blood, semen, and vaginal fluid are believed to contain sufficient viruses for transmission. Thus, sexual contact, injection of blood or blood products, and birth to an infected mother are well-established modes of transmission. Casual contact with infected individuals has not been found to transmit HIV.

Sexual transmission presently is the predominant mode of transmission. Receptive anal intercourse is more effective than other forms of sexual activity in transmitting HIV in homosexual men. Vaginal intercourse is largely responsible for transmission from men to women and from women to men. Intravenous inoculation of infected blood accounts for transmission of the virus among intravenous drug abusers who share needles. Inoculation of blood or blood products such as Factor VIII or XI concentrates has transmission infection in patients who have received such products and have not engaged in other risky activities. With current blood screening methods, the risk of transmitting HIV by a blood transfusion is estimated at between 1:40,000 and 1:225,000. Approximately 50% of babies born to infected mothers appear to develop HIV infections. Perinatal transmission occurs in utero or during delivery. Breast-feeding possibly transmits the virus as well.

Pathogenesis AIDS results from the infection and subsequent destruction of T4-lymphocytes by HIV. T4 lymphocytes play a key role in maintaining cellular immunity; their depletion leads to a multitude of abnormalities, which collectively undermine the immune response to infections. The infections often are lethal. The virus also infects other cells and promotes the development of certain tumors.

The activities and features of the viruses and T4 lymphocytes are central to understanding the pathogenesis of AIDS. HIV penetrates cells that contain CD4 receptors. Within the cytoplasm of the cell, the reverse transcriptase of the virus produces a DNA copy of its RNA genetic information. The DNA copy then is integrated into the genome of the host cell. A latency period ensues, after which immune activation results in viral replication and release from the cell, a process that destroys it.

T4 lymphocytes are the main target of HIV. T4 lymphocytes are responsible for inducing nearly every aspect of the immune response, including cytotoxic cells, suppressor cells, macrophages, B cells, natural killer cells, and even bone-marrow progenitor cells. Thus, replication of the virus leads in turn to depletion of the T4 lymphocytes and impairment of a multitude of immune responses.

Other cells with CD4 receptors also may be infected, including monocytes and macrophages. The monocytes are important in the pathogenesis of AIDS. Unlike T4 lymphocytes, the virus does not kill them rapidly. The monocytes thus harbor HIV and disseminate it to brain, bone marrow, and other organs.

Most of the clinical manifestations of AIDS result from opportunistic infections such as *Pneumocystis carinii* pneumonia, toxoplasmosis, cryptococcal meningitis, disseminated mycobacterium avium intracellulare cytomegalovirus, candida esophagitis, and several others. Other clinical manifestations result from the release of cytotoxins and growth factors from infected cells. Dementia in AIDS patients is fostered or caused by cytokines released from HIV-infected macrophages or monocytes rather than by HIV infection of neurons. Similarly, Kaposi's sarcoma appears to be due to the release of tumor-promoting factors from infected cells.

Symptoms and Signs—Infection by HIV usually is followed in a few days by an illness lasting 2 to 3 weeks. Symptoms include malaise, fever, weakness, rash, myalgia, and headache. The patient is then asymptomatic for several months or even several years. During this period antibodies to HIV can be detected in nearly all patients, but the virus and the clinical picture are in a period of latency. When HIV is activated and replicates, the number of T4 lymphocytes declines, and symptoms and signs begin to appear. Over 5 to 10 years after infection, 25–50% of persons will progress to overt AIDS without treatment. Most patients first experience fatigue, anorexia, weight loss, and unexplained fever. Chronic lymph node enlargement, particularly in the neck, is common. Diarrhea often ensues. Nonproductive cough and dyspnea often herald the presence of opportunistic pneumonia. A host of neuropsychiatric symptoms may occur, including confusion, headache, seizures, focal weakness, personality changes, and impaired memory. This is an abbreviated list since every organ may be involved. The possible clinical expressions are vast.

Diagnosis—The diagnosis of HIV infection is made by the detection of the HIV virus using any of the following methods: detecting antibodies to the virus, detecting the viral p24 antigen, detecting viral nucleic acid, or culturing the virus from tissue or blood. The most widely used of these methods is the serology test for antibodies. Antibodies to HIV are first detectable 6 to 12 weeks after infection though may be delayed as much as 6 months.

AIDS is a clinical definition. In 1993 the criteria for AIDS was re-defined. Patients are now classified as having AIDS if they have any of several clinical diseases known as "AIDS indicator conditions" and/or a CD4 count of less than 200/mm^3. AIDS indicator conditions are mostly opportunistic or recurrent infections that have become associated with advanced HIV disease.

BIBLIOGRAPHY

Mandell GL, Douglas RG, Bennett JE, eds. Mandell, Douglas, and Bennett's Principles and Practices of Infectious Diseases, 5th ed. Philadelphia: Churchill Livingstone, 2000.

Callen JP et al, eds. Dermatological Signs of Internal Disease, 3rd ed. Philadelphia: WB Saunders, 2002.

Kelley WN, et al, eds. Textbook of Internal Medicine, 3rd ed. Philadelphia: Lippincott-Raven, 1997.

Felig P, Frohman LA, eds. Endocrinology and Metabolism, 4th ed. New York: McGraw-Hill, 2001.

Klippel JH, Weyand CM, Crofford LJ, et al, eds. Primer on the Rheumatic Diseases, 12th ed. Atlanta: Arthritis Foundation, 2001.

Kjeldsberg CR, ed. Practical Diagnosis of Hematologic Disorders, 3rd ed. Chicago: ASCP Press, 2000.

George RB, et al, eds. Chest Medicine: Essentials of Pulmonary and Critical Care medicine, 4th ed. Philadelphia: Lippincott Williams and Wilkins, 2000.

CHAPTER 57

Drug Absorption, Action, and Disposition

Michael R Franklin, PhD
Donald N Franz, PhD

Although drugs differ widely in their pharmacodynamic effects and clinical applications; in penetration, absorption, and usual route of administration; in distribution among the body tissues; and in disposition and mode of termination of action, there are certain general principles that help explain these differences. These principles have both pharmaceutic and therapeutic implications. They facilitate an understanding of both the features that are common to a class of drugs and the differences among the members of that class.

For a drug to act it must be absorbed, transported to the appropriate tissue or organ, penetrate to the responding cell surface or subcellular structure, and elicit a response or alter ongoing processes. The drug may be distributed simultaneously or sequentially to a number of tissues, bound or stored, metabolized to inactive or active products, or excreted. The history of a drug in the body is summarized in Figure 57-1. Each of the processes or events depicted relates importantly to therapeutic and toxic effects of a drug and to the mode of administration, and drug design must take each into account. Since the effect elicited by a drug is its *raison d'être, drug action,* and *effect* are discussed first in the text that follows, even though they are preceded by other events.

DRUG ACTION AND EFFECT

The word *drug* imposes an action-effect context within which the properties of a substance are described. The description of necessity must include the pertinent properties of the recipient of the drug. Thus, when a drug is defined as an analgesic, it is implied that the recipient reacts to a noxious stimulus in a certain way, called pain. (Studies indicate that pain is not simply the *perception* of a certain kind of stimulus but rather, a *reaction* to the perception of a variety of kinds of stimuli or stimulus patterns.) Both because the pertinent properties are locked into the complex and somewhat imprecise biological context and because the types of possible response are many, descriptions of the properties of drugs tend to emphasize the qualitative features of the effects they elicit. Thus, a drug may be described as having analgesic, vasodepressor, convulsant, antibacterial, etc, properties. The specific effect (or use) categories into which the many drugs may be placed are the subject of Chapters 64 through 89 and are not elaborated upon in this chapter. However, the description of a drug does not end with the enumeration of the responses it may elicit. There are certain intrinsic properties of the drug-recipient system that can be described in quantitative terms and that are essential to the full description of the drug and to the validation of the drug for specific uses. Under *Definitions and Concepts* below, certain general terms are defined in qualitative language; under *Dose-Effect Relationships,* the foundation is laid for an appreciation of some of the quantitative aspects of pharmacodynamics.

DEFINITIONS AND CONCEPTS

In the field of pharmacology, the vocabulary that is unique to the discipline is relatively small, and the general vocabulary is that of the biological sciences and chemistry. Nevertheless, there are a few definitions that are important to the proper understanding of pharmacology. It is necessary to differentiate among action, effect, selectivity, dose, potency, and efficacy.

ACTION VS EFFECT—The *effect* of a drug is an *alteration of function* of the structure or process upon which the drug acts. It is common to use the term *action* as a synonym for effect. However, action precedes effect. *Action* is the *alteration of condition* that brings about the effect.

The final effect of a drug may be far removed from its site of action. For example, the diuresis subsequent to the ingestion of ethanol does not result from an action on the kidney but instead from a depression of activity in the region of the hypothalamus, which regulates the release of antidiuretic hormone from the posterior pituitary gland. The alteration of hypothalamic function is, of course, also an effect of the drug, as is each subsequent change in the chain of events leading to diuresis. The action of ethanol was exerted only at the initial step, each subsequent effect being then the action to a following step.

MULTIPLE EFFECTS—No known drug is capable of exerting a single effect, although a number are known that appear to have a single mechanism of action. Multiple effects may derive from a single mechanism of action. For example, the inhibition of acetylcholinesterase by physostigmine will elicit an effect at every site where acetylcholine is produced, is potentially active, and is hydrolyzed by cholinesterase. Thus, physostigmine elicits a constellation of effects.

A drug also can cause multiple effects at several different sites by a single action at only one site, providing that the function initially altered at the site of action ramifies to control other functions at distant sites. Thus, a drug that suppresses steroid synthesis in the liver may not only lower serum cholesterol, impair nerve myelination and function, and alter the condition of the skin (as a consequence of cholesterol deficiency) but also may affect digestive functions (because of a deficiency in bile acids) and alter adrenocortical and sexual hormonal balance.

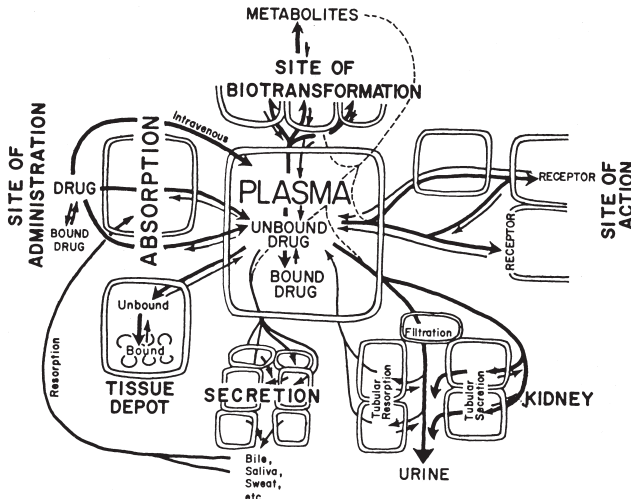


Figure 57-1. The absorption, distribution, action, and elimination of a drug (arrows represent drug movement). Intravenous administration is the only process by which a drug may enter a compartment without passing through a biological membrane. Note that drugs excreted in bile and saliva may be resorbed.

Although a single action can give rise to multiple effects, most drugs exert multiple actions. The various actions may be related, as for example, the sympathomimetic effects of phenylephrine that accrue to its structural similarity to norepinephrine and its ability to exert sympathetic responses, or the actions may be unrelated, as with the actions of morphine to interfere with the release of acetylcholine from certain autonomic nerves, block some actions of 5-hydroxytryptamine (serotonin), and release histamine. Many drugs bring about immunological (allergic or hypersensitivity) responses that bear no relation to the other pharmacodynamic actions of the drug.

SELECTIVITY—Despite the potential most drugs have for eliciting multiple effects, one effect is generally more readily elicitable than another. This differential responsiveness is called *selectivity*. It usually is considered to be a property of the drug, but it is also a property of the constitution and biodynamics of the recipient subject or patient.

Selectivity may come about in several ways. The subcellular structure (receptor) with which a drug combines to initiate one response may have a higher affinity for the drug than that for some other action. Atropine, for example, has a much higher affinity for muscarinic receptors that subserve the function of sweating than it does for the nicotinic receptors that subserve voluntary neuromuscular transmission, so that suppression of sweating can be achieved with only a tiny fraction of the dose necessary to cause paralysis of the skeletal muscles. A drug may be distributed unevenly, so that it reaches a higher concentration at one site than throughout the tissues generally; chloroquine is much more effective against hepatic than intestinal (colonic) amebiasis because it reaches a much higher concentration in the liver than in the wall of the colon. An affected function may be much more critical to, or have less reserve in, one organ than in another, so that a drug will be predisposed to elicit an effect at the more critical site. Some inhibitors of dopa decarboxylase (which is also 5-hydroxytryptophan decarboxylase) depress the synthesis of histamine more than that of either norepinephrine or 5- hydroxytryptamine (serotonin), even though histidine decarboxylase is less sensitive to the drug, simply because histidine decarboxylase is the only step and, hence, is rate-limiting in the biosynthesis of histamine. Dopa decarboxylase is not rate limiting in the synthesis of either norepinephrine or 5-hydroxytryptamine until the enzyme is nearly completely inhibited. Another example of the determination of selectivity by the critical balance of the affected function is that of the mercurial diuretic drugs. An inhibition of only 1% in the tubular resorption of glomerular filtrate usually will double urine flow, since 99% of the glomerular filtrate is normally resorbed. Aside from the question of the possible concentration of diuretics in the urine, a drug-induced reduction of 1% in sulfhydryl enzyme activity in tissues other than the kidney usually is not accompanied by an observable change in function. Selectivity also can be determined by the pattern of distribution of inactivating or activating enzymes among the tissues and by other factors.

DOSE—Even the uninitiated person knows that the dose of a drug is the amount administered. However, the appropriate dose of a drug is not some unvarying quantity, a fact sometimes overlooked by pharmacists, official committees, and physicians. The practice of pharmacy is entrapped in a system of fixed-dose formulations, so that fine adjustments in dosage are often difficult to achieve. Fortunately, there is usually a rather wide latitude allowable in dosages. It is obvious that the size of the recipient individual should have a bearing upon the dose, and the physician may elect to administer the drug on a body-weight or surface-area basis rather than as a fixed dose. Usually, however, a fixed dose is given to all adults, unless the adult is exceptionally large or small. The dose for infants and children often is determined by one of several formulas that take into account age or weight, depending on the age group of the child and the type of action exerted by the drug. Infants, relatively, are more sensitive to many drugs, often because systems involved in the inactivation and elimination of the drugs may not be developed fully in the infant.

The nutritional condition of the patient, the mental outlook, the presence of pain or discomfort, the severity of the condition being treated, the presence of secondary disease or pathology, and genetic and many other factors affect the dose of a drug necessary to achieve a given therapeutic response or to cause an untoward effect (Chapter 61). Even two apparently well-matched normal persons may require widely different doses for the same intensity of effect. Furthermore, a drug is not always employed for the same effect and, hence, not in the same dose. For example, the dose of a progestin necessary for an oral contraceptive effect is considerably different from that necessary to prevent spontaneous abortion, and a dose of an estrogen for the treatment of the menopause is much too small for the treatment of prostatic carcinoma.

From the above, it is evident that the wise physician knows that *the dose of a drug* is not a rigid quantity but rather that which is necessary and can be tolerated and individualizes the regimen accordingly. The wise pharmacist also recognizes that official or manufacturer's recommended doses are sometimes quite narrowly defined and should serve only as a useful guide rather than as an imperative.

POTENCY AND EFFICACY—The *potency* of a drug is the reciprocal of dose. Thus, it will have the units of persons/unit weight of drug or body weight/unit weight of drug, etc. Potency generally has little utility other than to provide a means of comparing the relative activities of drugs in a series, in which case *relative potency,* relative to some prototypic member of the series, is a parameter commonly used among pharmacologists and in the pharmaceutical industry.

Whether a given drug is more potent than another has little bearing on its clinical usefulness, provided that the potency is not so low that the size of the dose is physically unmanageable or the cost of treatment is higher than with an equivalent drug. If a drug is less potent but more selective, it is the one to be preferred. Promotional arguments in favor of a more potent drug thus are irrelevant to the important considerations that should govern the choice of a drug. However, it sometimes occurs that drugs of the same class differ in the maximum intensity of effect; that is, some drugs of the class may be less efficacious than others, irrespective of how large a dose is used.

Efficacy connotes the property of a drug to achieve the desired response, and *maximum efficacy* denotes the maximum achievable effect. Even huge doses of codeine often cannot achieve the relief from severe pain that relatively small doses

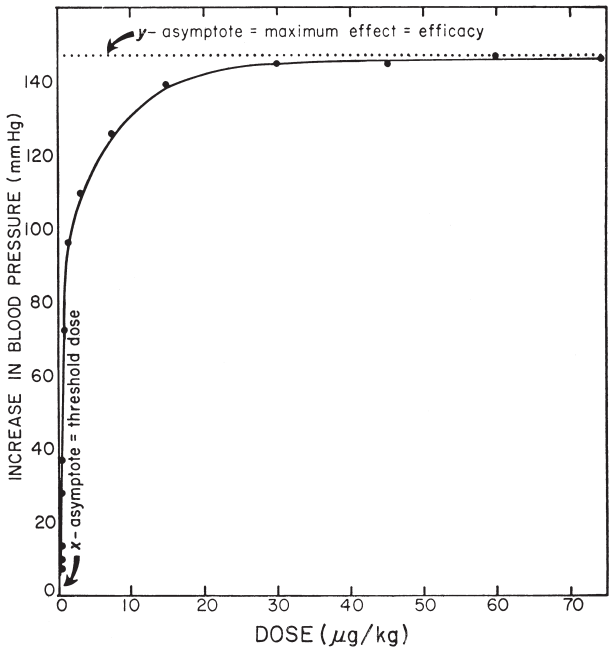


Figure 57-2. The relationship of the intensity of the blood-pressure response of the cat to the intravenous dose of norepinephrine.

of morphine can; thus, codeine is said to have a lower maximum efficacy than morphine. Efficacy is one of the primary determinants of the choice of a drug.

DOSE-EFFECT RELATIONSHIPS

The importance of knowing how changes in the intensity of response to a drug vary with the dose is virtually self-evident. Both the physician, who prescribes or administers a drug, and the manufacturer, who must package the drug in appropriate dose sizes, must translate such knowledge into everyday practice. Theoretical or molecular pharmacologists also study such relationships in inquiries into mechanism of action and receptor theory It is necessary to define two types of relationships: (1) dose-intensity relationship, ie, the manner in which the intensity of effect in the individual recipient relates to dose, and (2) dose-frequency relationship, ie, the manner in which the number of responders among a population of recipients relates to dose.

DOSE–INTENSITY OF EFFECT RELATIONSHIPS— Whether the intensity of effect is determined *in vivo* (eg, the blood-pressure response to epinephrine in the human patient) or *in vitro* (eg, the response of the isolated guinea pig ileum to histamine), the dose–intensity of effect (often called dose-effect) curve usually has a characteristic shape, namely a curve that closely resembles one quadrant of a rectangular hyperbola.

In the dose-intensity curve depicted in Figure 57-2, the curve appears to intercept the x axis at 0 only because the lower doses are quite small on the scale of the abscissa, the smallest dose being 1.5×10^{-3} μg. Actually, the x intercept has a positive value, since a finite dose of drug is required to bring about a response, this lowest effective dose being known as the *threshold dose*. Statistics and chemical kinetics predict that the curve should approach the y axis asymptotically. However, if the intensity of the measured variable does not start from zero, the curve possibly may have a positive y intercept (or negative x intercept), especially if the ongoing basal activity before the drug is given is closely related to that induced by the drug.

In practice, instead of an asymptote to the y axis, dose-intensity curves nearly always show an upward concave foot at the origin of the curve, so that the curve has a lopsided sigmoid shape. At high doses, the curve approaches an asymptote that is parallel to the x axis, and the value of the asymptote establishes the maximum possible response to the drug, or *maximum efficacy*. However, experimental data in the regions of the asymptotes generally are too erratic to permit an exact definition of the curve at very low and very high doses. The example shown represents an unusually good set of data.

Because the dose range may be 100- or 1000-fold from the lowest to the highest dose, it has become the practice to plot dose-intensity curves on a logarithmic scale of abscissa (ie, to plot the log of dose versus the intensity of effect). Figure 57-3 is such a semilogarithmic plot of the same data used in Figure 57-2. In the figure the intensity of effect is plotted both in absolute units (at the left) or in relative units, as percentages (at the right).

Although no new information is created by a semilogarithmic representation, the curve is stretched in such a way as to facilitate the inspection of the data; the comparison of results

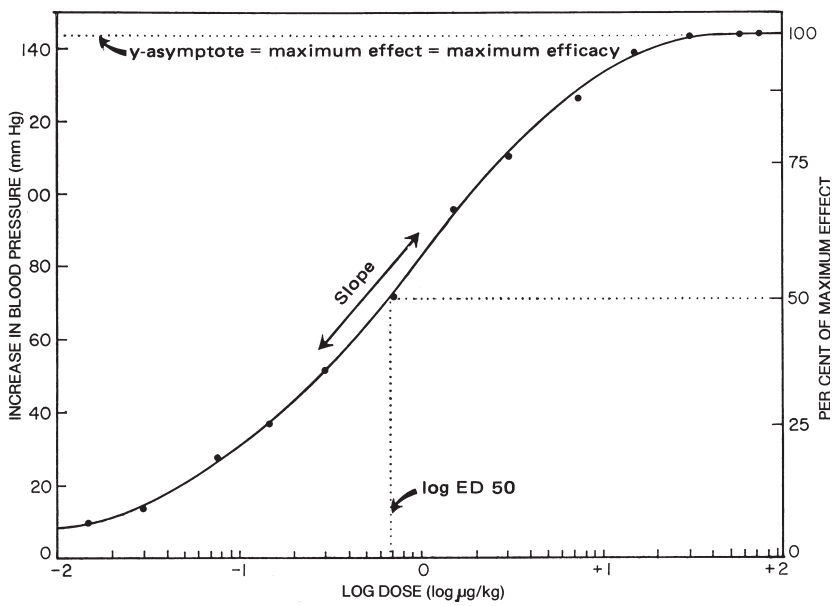


Figure 57-3. The relationship of the intensity of the blood-pressure response of the cat to the log of the intravenous dose of norepinephrine.

from multiple observations and the testing of different drugs also is rendered easier. In the example shown, the curve is essentially what is called a *sigmoid curve* and is nearly symmetrical about the point that represents an intensity equal to 50% of the maximal effect (ie, about the midpoint). The symmetry follows from the rectangular hyperbolic character of the previous Cartesian plot (see Fig 57-2). The semilogarithmic plot reveals better the dose-effect relationships in the low-dose range, which are lost in the steep slope of the Cartesian plot. Furthermore, the data about the midpoint are almost a straight line; the nearly linear portion covers approximately 50% of the curve. The slope of the linear portion of the curve or, more correctly, the slope at the point of inflection, has theoretical significance (see *Drug Receptors and Receptor Theory*).

The upper portion of the curve approaches an asymptote, which is the same as that in the Cartesian plot. If the response system is completely at rest before the drug is administered, the lower portion of the curve should be asymptotic to the *x* axis. Both asymptotes and the symmetry derive from the law of mass action.

Dose-intensity curves often deviate from the ideal configuration illustrated and discussed above. Usually, the deviate curve remains sigmoid but not extended symmetrically about the midpoint of the *linear* segment. Occasionally, other shapes occur. Deviations may derive from multiple actions that converge upon the same final effector system, from varying degrees of metabolic alteration of the drug at different doses, from modulation of the response by feedback systems, from nonlinearity in the relationship between action and effect, or from other causes.

It is frequently necessary to identify the dose that elicits a given intensity of effect. The intensity of effect that is generally designated is 50% of maximum intensity. The corresponding dose is called the *50% effective dose,* or *individual ED50* (see Fig 57-3). The use of the adjective *individual* distinguishes the ED50 based upon the intensity of effect from the median effective dose, also abbreviated ED50, determined from frequency of response data in a population (see *Dose-Frequency Relationships,* this chapter).

Drugs that elicit the same quality of effect may be compared graphically. In Figure 57-4, five hypothetical drugs are compared. Drugs *A, B, C,* and *E* all can achieve the same maximum effect, which suggests that the same effector system may be common to all. *D* possibly may be working through the same effector system, but there are no *a priori* reasons to think this is so. Only *A* and *B* have parallel curves and common slopes. Common slopes are consistent with, but in no way prove, the idea that *A* and *B* not only act through the same effector system but also by the same mechanism. Although drug-receptor theory (see *Drug Receptors and Receptor Theory*) requires that the curves of identical mechanism have equal slopes, examples of exceptions are known. Furthermore, mass-law statistics require that all simple drug-receptor interactions generate the same slope; only when slopes depart from this universal slope in accordance with distinctive characteristics of the response system do they provide evidence of specific mechanisms.

The relative potency of any drug may be obtained by dividing the ED50 of the standard, or prototypic, drug by that of the drug in question. Any level of effect other than 50% may be used, but it should be recognized that when the slopes are not parallel, the relative potency depends upon the intensity of effect chosen. Thus, the potency of *A* relative to *C* (see Fig 57-4) calculated from the ED50 will be smaller than that calculated from the ED25.

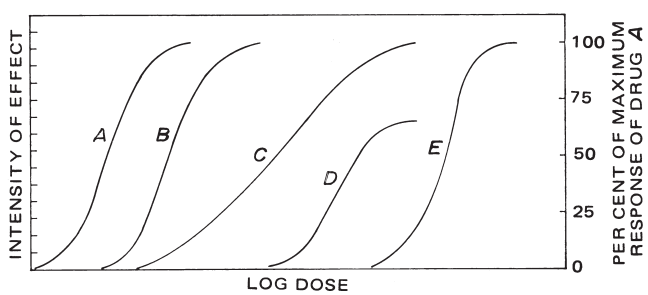


Figure 57-4. Log dose–intensity of effect curves of five different hypothetical drugs (see text for explanation).

The low maximum intensity inducible by *D* poses even more complications in the determination of relative potency than do the unequal slopes of the other drugs. If its dose-intensity curve is plotted in terms of percentage of its own maximum effect, its relative inefficacy is obscured, and the limitations of relative potency at the ED50 level will not be evident. This dilemma underscores the fact that drugs can be compared better from their entire dose-intensity curves than from a single derived number like ED50 or relative potency.

Drugs that elicit multiple effects will generate a dose-intensity curve for each effect. Even though the various effects may be qualitatively different, the several curves may be plotted together on a common scale of abscissa, and the intensity may be expressed in terms of percentage of maximum effect; thus, all curves can share a common scale of ordinates in addition to a common abscissa. Separate scales of ordinates could be employed, but this would make it harder to compare data.

The selectivity of a drug can be determined by noting what percentage of maximum of one effect can be achieved before a second effect occurs. As with relative potency, selectivity may be expressed in terms of the ratio between the ED50 for one effect and that for another effect, or a ratio at some other intensity of effect. As with relative potency, difficulties follow from nonparallelism. In such instances, selectivity expressed in dose ratios varies from one intensity level to another.

When the dose-intensity curves for a number of subjects are compared, it is found that they vary considerably from individual to individual in many respects; eg, threshold dose, midpoint, maximum intensity, and sometimes even slope. By averaging the intensities of the effect at each dose, an average dose-intensity curve can be constructed.

Average dose-intensity curves enjoy a limited application in comparing drugs. A single line expressing an average response has little value in predicting individual responses unless it is accompanied by some expression of the range of the effect at the various doses. This may be done by indicating the standard error of the response at each dose. Occasionally, a simple scatter diagram is plotted in lieu of an average curve and statistical parameters. An average dose-intensity curve also may be constructed from a population in which different individuals receive different doses; if sufficiently large populations are employed, the average curves determined by the two methods will approximate each other.

It is obvious that the determination of such average curves from a population large enough to be statistically meaningful requires a great deal of work. Retrospective clinical data occasionally are treated in this way, but prospective studies infrequently are designed in advance to yield average curves. The usual practice in comparing drugs is to employ a quantal (all-or-none) endpoint and plot the frequency or cumulative frequency of response over the dose range, as discussed below.

DOSE–FREQUENCY OF RESPONSE RELATIONSHIPS—When an endpoint is truly all-or-none, such as death, it is an easy matter to plot the number of responding individuals (eg, dead subjects) at each dose of drug or intoxicant. Many other responses that vary in intensity can be treated as all-or-none if simply the presence or absence of a response (eg, cough or no cough, convulsion or no convulsion) is recorded, without regard to the intensity of the response when it occurs. When the response changes from the basal or control state in a less abrupt manner (eg, tachycardia, miosis, rate of gastric secretion), it may be necessary to designate arbitrarily some particular intensity of effect as the endpoint. If the endpoint is taken as an increase in heart rate of 20 beats/min, all individuals whose tachycardia is less than 20 beats/min would be recorded as nonresponders, while all those with 20 or above would be recorded as responders. When the percentage of responders in the population is plotted against the dose, a characteristic dose-response curve, more

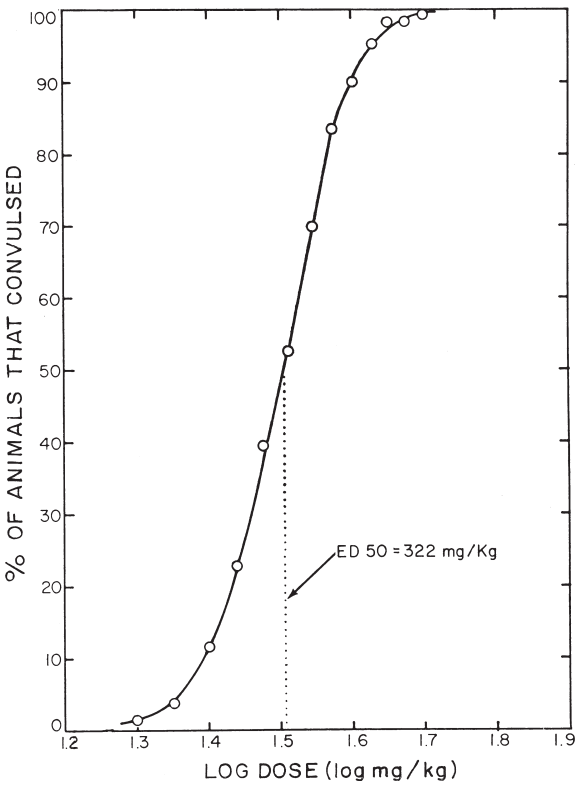


Figure 57-5. The relationship of the number of responders in a population of mice to the dose of pentylenetetrazole.

properly called a *dose–cumulative frequency* or *dose-percentage* curve, is generated. Such a curve is, in fact, a cumulative frequency-distribution curve, the percentage of responders at a given dose being the frequency of response.

Dose–cumulative frequency curves are generally of the same geometric shape as dose-intensity curves (namely, sigmoid) when frequency is plotted against log dose (Fig 57-5). The tendency of the cumulated frequency of response (ie, percentage) to be linearly proportional to the log of the dose in the middle of the dose range is called the *Weber-Fechner law,* although it is not invariable, as a true natural law should be. In many instances, the cumulative frequency is simply proportional to dose rather than log dose. The Weber-Fechner law applies to either dose-intensity or dose–cumulative frequency data. The similarity between dose-frequency and dose-intensity curves may be more than fortuitous, since the intensity of response will usually have an approximately linear relationship to the percentage of responding *units* (smooth muscle cells, nerve fibers, etc) and, hence, is also a type of cumulative frequency of response. These are the same kind of statistics that govern the law of mass action.

If only the increase in the number of responders with each new dose is plotted, instead of the cumulative percentage of responders, a bell-shaped curve is obtained. This curve is the first derivative of the dose–cumulative frequency curve and is a *frequency-distribution* curve. The distribution will be symmetrical—ie, *normal* or Gaussian (see Chapter 12)—only if the dose–cumulative frequency curve is symmetrically hyperbolic. Because most dose–cumulative frequency curves are more nearly symmetrical when plotted semilogarithmically (ie, as log dose), dose–cumulative frequency curves are usually *log-normal*.

Since the dose-intensity and dose–cumulative frequency curves are basically similar in shape, it follows that the curves have similar defining characteristics, such as ED50, maximum effect (maximum efficacy), and slope. In dose–cumulative frequency data, the ED50 *(median effective dose)* is the dose to which 50% of the population responds (see Fig 57-5). If the frequency distribution is normal, the ED50 is both the arithmetic mean and the median dose and is represented by the midpoint on the curve; if the distribution is log-normal, the ED50 is the median dose but not the arithmetic mean dose. The efficacy is the cumulative frequency summed over all doses; it is usually, but not always, 100%. The slope is characteristic of both the drug and the test population. Even two drugs of identical mechanism may give rise to different slopes in dose-percentage curves, whereas in dose-intensity curves the slopes are the same.

Statistical parameters (such as standard deviation), in addition to ED50, maximum cumulative frequency (efficacy) and slope, characterize dose–cumulative frequency relationships (see Chapter 12).

There are several formulations for dose–cumulative frequency curves, some of which are employed only to define the linear segment of a curve and to determine the statistical parameters of this segment. For the statistical treatment of dose-frequency data, see Chapter 12. One simple mathematical expression of the entire log-symmetrical sigmoid curve is

$$\log \text{dose} = \text{K} + f \log \left(\frac{\%\ \text{response}}{100\% - \text{response}} \right) \qquad (1)$$

where percentage response may be either the percentage of maximum intensity or the percentage of a population responding. The equation is thus basically the same for both log normal dose-intensity and log normal dose-percentage relationships. K is a constant that is characteristic of the midpoint of the curve, or ED50, and $1/f$ is characteristically related to the slope of the linear segment, which, in turn is closely related to the standard deviation of the derivative log-normal frequency-distribution curve.

The comparison of dose-percentage relationships among drugs is subject to the pitfalls indicated for dose-intensity comparisons, namely, that when the slopes of the curves are not the same (ie, the dose-percentage curves are not parallel), it is necessary to state at which level of response a potency ratio is calculated. As with dose-intensity data, potencies generally are calculated from the ED50, but potency ratios may be calculated for any arbitrary percentage response. The expression of selectivity is, likewise, subject to similar qualifications, inasmuch as the dose-percentage curves for the several effects are usually nonparallel.

The term *therapeutic index* is used to designate a quantitative statement of the selectivity of a drug when a therapeutic and an untoward effect are being compared. If the untoward effect is designated T (for toxic) and the therapeutic effect, E, the therapeutic index may be defined as TD50/ED50 or a similar ratio at some other arbitrary levels of response. The TD and the ED are not required to express the same percentage of response; some clinicians use the ratio TD1/ED99 or TD5/ED95, based on the rationale that if the untoward effect is serious, it is important to use a most-severe therapeutic index in passing judgment upon the drug. Unfortunately, therapeutic indices are known in man for only a few drugs.

There will be a different therapeutic index for each untoward effect that a drug may elicit and, if there is more than one therapeutic effect, a family of therapeutic indices for each therapeutic effect. However, in clinical practice, it is customary to distinguish among the various toxicities by indicating the percentage incidence of a given side effect.

VARIATIONS IN RESPONSE AND RESPONSIVENESS—From the above discussion of dose-frequency relationships and Chapter 12, it is obvious that in a normal population of persons there may be quite a large difference in the dose required to elicit a given response in the least-responsive member of the population and that to elicit the response in the most-responsive member. The difference ordinarily will be a function of the slope of the dose-percentage curve or, in statistical terms, of the standard deviation. If the standard deviation is large, the extremes of responsiveness of responders are likewise large.

In a normal population, 95.46% of the population responds to doses within two standard deviations from the ED50 and

99.73% within three standard deviations. In log-normal populations, the same distribution applies when standard deviation is expressed as log dose.

In the population represented in Figure 57-5, 2.25% of the population (two standard deviations from the median) would require a dose more than 1.4 times the ED50; an equally small percentage would respond to 0.7 of the ED50. The physician who is unfamiliar with statistics is apt to consider the 2.25% at either extreme to be abnormal reactors. The statistician will argue that these 4.5% are within the normal population and that only those who respond well outside the normal population, at least three standard deviations from the median, deserve to be called abnormal.

Irrespective of whether the criteria of abnormality that the physician or the statistician obtain, the term *hyporeactive* applies to those individuals who require abnormally high doses and *hyperreactive* to those who require abnormally low doses. The terms *hyporesponsive* and *hyperresponsive* also may be used. It is incorrect to use the terms hyposensitive and hypersensitive in this context; *hypersensitivity* denotes an allergic response to a drug and should not be used to refer to hyperreactivity. The term *supersensitivity* correctly applies to hyperreactivity that results from denervation of the effector organ; it is often more definitively called denervation supersensitivity. Sometimes hyporeactivity is the result of an immunochemical deactivation of the drug, or *immunity*. Hyporeactivity should be distinguished from an increased dose requirement that results from a severe pathological condition. Severe pain requires large doses of analgesics, but the patient is not a hyporeactor; what has changed is the baseline from which the endpoint quantum is measured. The responsiveness of a patient to certain drugs sometimes may be determined by the history of previous exposure to appropriate drugs.

Tolerance is a diminution in responsiveness as use of the drug continues. The consequence of tolerance is an increase in the dose requirement. It may be due to an increase in the rate of elimination of drug (as discussed elsewhere in this chapter), to reflex or other compensatory homeostatic adjustments, to a decrease in the number of receptors or in the number of enzyme molecules or other coupling proteins in the effector sequence, to exhaustion of the effector system or depletion of mediators, to the development of immunity, or to other mechanisms. Tolerance may be gradual, requiring many doses and days to months to develop, or acute, requiring only the first or a few doses and only minutes to hours to develop. Acute tolerance is called *tachyphylaxis*.

Drug resistance is the decrease in responsiveness of microorganisms, neoplasms, or pests to chemotherapeutic agents, antineoplastics, or pesticides, respectively. It is not tolerance in the sense that the sensitivity of the individual microorganism or cancer cell decreases; rather, it is the survival of normally unresponsive cells, which then pass the genetic factors of resistance on to their progeny.

Patients who fail to respond to a drug are called *refractory*. Refractoriness may result from tolerance or resistance, but it also may result from the progression of pathological states that negate the response or render the response incapable of surmounting an overwhelming pathology. Rarely, it may result from a poorly developed receptor or response system.

Sometimes a drug evokes an unusual response that is *qualitatively* different from the expected response. Such an unexpected response is called a *meta-reaction*. A not uncommon *meta*-reaction is a central nervous system (CNS) stimulant rather than depressant effect of phenobarbital, especially in women. Pain and certain pathological states sometimes favor *meta*-reactivity. Responses that are different in infants or the aged from those in young and middle-aged people are not *meta*-reactions if the response is usual in the age group. The term *idiosyncrasy* also denotes *meta*-reactivity, but the word has been so abused that it is recommended that it be dropped. Although hypersensitivity may cause unusual effects, it is not included in *meta*-reactivity.

DRUG RECEPTORS AND RECEPTOR THEORY

Most drugs act by combining with some key substance in the biological milieu that has an important regulatory function in the target organ or tissue. This biological partner of the drug goes by the name *receptive substance* or *drug receptor*. The receptive substance is considered mostly to be a cellular constituent, although in a few instances it may be extracellular, as the cholinesterases are, in part. The receptive substance is thought of as having a special chemical affinity and structural requirements for the drug. Drugs such as emollients, which have a physical rather than chemical basis for their action, obviously do not act upon receptors. Drugs such as demulcents and astringents, which act in a nonselective or nonspecific chemical way, also are not considered to act upon receptors, since the candidate receptors have neither sharp chemical nor biological definition. Even antacids, which react with the extremely well defined hydronium ion, cannot be said to have a receptor, since the reactive proton has no permanent biological residence.

Because of early preoccupation with physical theories of action and the classical and illogical dichotomy of chemical and physical molecular interaction, there is a reluctance to admit receptors for drugs such as general anesthetics, certain electrolytes, etc, which generally are not accepted to combine selectively with distinct cellular or organelle membrane constituents. The word receptor often is used inconsistently and intuitively. However, the term is a legitimate symbol for that biological structure with which a drug interacts to initiate a response. Ignorance of the identities of many receptors does not detract from, but rather increases, the importance of the term and general concept.

Once a receptor is identified, it frequently is no longer thought of as a receptor, although such identification may afford the basis of profound advances in receptor theory. Since the effects of anticholinesterases are derived only indirectly from inhibition of cholinesterase and no drugs are known that stimulate the enzyme, it may be argued that it is not a receptor. Nevertheless, a number of drugs ultimately act indirectly through the inhibition of such modulator enzymes, and it is important for the theoretician to develop models based upon such indirect interrelations.

Enzymes, of course, readily suggest themselves as candidates for receptors. However, there is more to cellular function than enzymes. Receptors may be membrane or intracellular constituents that govern the spatial orientation of enzymes, gene expression, compartmentalization of the cytoplasm, contractile or compliant properties of subcellular structures, or permeability and electrical properties of membranes. For nearly every cellular constituent there can be imagined a possible way for a drug to affect its function; therefore, few cellular constituents can be dismissed *a priori* as possible receptors. All the receptors for neurotransmitters and autonomic agonists are membrane proteins with agonist-binding groups projecting into the extracellular space. The transducing apparatus, whereby an occupied receptor elicits a response, is called a *coupling system*. Excitatory neurotransmitters in the CNS, and ACh receptors elsewhere, are coupled to ion channels that, when opened, permit the rapid ingress, especially of sodium ions. Each ion channel is composed of five subunits, and each subunit has four transmembrane, spanning regions. GABA (γ-aminobutyric acid) and glycine are coupled to inhibitory chloride channels. Each of these receptors is composed of pentameric proteins, each of which has two to four different types of subunits. Benzodiazepine receptors are coupled to the GABA-receptor. Beta-adrenergic receptors, histamine (H2) receptors, and a number of receptors for polypeptide hormones interact with a stimula-

tory GDP/GTP-binding protein (G-protein) that can activate the enzyme adenylate cyclase. The cyclase then produces 3′,5′-cyclic AMP (cAMP), which, in turn, activates protein kinases. Other receptors interact with inhibitory G-proteins. Some receptors couple to guanylate cyclase.

Alpha-adrenergic α_1, some muscarinic (M_1 and M_3), and various other receptors couple to the membrane enzyme, phospholipase-C, which cleaves inositol phosphates from phosphoinositides. The cleavage product, 1,4,5-inositol triphosphate (IP_3), then causes an increase in intracellular calcium, whereas the product, diacylglycerol (DAG), activates kinase-C. There are a number of other less ubiquitous coupling systems. Substances such as cAMP, cGMP, IP_3, and DAG are called *second messengers*.

It has been found that there may be several different receptors for a given agonist. Differences may be shown not only in the types of coupling systems and effects but also by differential binding of agonists and antagonists, desensitization kinetics, physical and chemical properties, genes and amino acid sequences. The differentiation among receptor subtypes is called *receptor classification*. Receptor subtypes are designated by Greek or Arabic alphabetical prefixes and/or numerical subscripts. There are at least two each of beta-adrenergic, histaminergic, serotoninergic, GABAergic, and benzodiazepine receptors; three each of muscarinic and alpha-adrenergic; and five of opioid receptor subtypes.

OCCUPATION AND OTHER THEORIES

Drug-receptor interactions are governed by the law of mass action. However, most chemical applications of mass law are concerned with the rate at which reagents disappear or products are formed, whereas receptor theory usually concerns itself with the fraction of the receptors combined with a drug. The usual concept is that only when the receptor actually is occupied by the drug is its function transformed in such a way as to elicit a response. This concept has become known as the *occupation theory*. The earliest clear statement of its assumptions and formulations is often credited to Clark in 1926, but both Langley and Hill made important contributions to the theory in the first two decades of the 20th century.

In all receptor theories, the terms agonist, partial agonist, and antagonist are employed. An *agonist* is a drug that combines with a receptor to initiate a response.

In the classical occupation theory, two attributes of the drug are required: (1) *affinity,* a measure of the equilibrium constant of the drug-receptor interaction, and (2) *intrinsic activity,* or *intrinsic efficacy* (not to be confused with efficacy as intensity of effect), a measure of the ability of the drug to induce a positive change in the function of the receptor.

A *partial agonist* is a drug that can elicit some but not a maximal effect and that antagonizes an agonist. In the occupation theory it would be a drug with a favorable affinity but a low intrinsic activity.

A *competitive antagonist* is a drug that occupies a significant proportion of the receptors and thereby preempts them from reacting maximally with an agonist. In the occupation theory the prerequisite property is affinity without intrinsic activity.

A *noncompetitive antagonist* may react with the receptor in such a way as not to prevent agonist-receptor combination but to prevent the combination from initiating a response, or it may act to inhibit some subsequent event in the chain of action-effect-action-effect that leads to the final overt response.

The mathematical formulation of the receptor theories derives directly from the law of mass action and chemical kinetics. Certain assumptions are required to simplify calculations. The key assumption is that the intensity of effect is a direct linear function of the proportion of receptors occupied. The correctness of this assumption is most improbable on the basis of theoretical considerations, but empirically it appears to be a close enough approximation to be useful. A second assumption upon which formulations are based is that the drug-receptor interaction is at equilibrium. Another common assumption is that the number of molecules of receptor is negligibly small compared with that of the drug. This assumption is undoubtedly true in most instances, and departures from this situation greatly complicate the mathematical expression of drug-receptor interactions.

The first clearly stated mathematical formulation of drug-receptor kinetics was that of Clark.[1] In his equation

$$Kx^n = \frac{y}{100 - y} \tag{2}$$

where K is the affinity constant, x is the concentration of drug, n is the molecularity of the reaction, and y is the percentage of maximum response. Clark assumed that y was a linear function of the percentage of receptors occupied by the drug, so that y could also symbolize the percentage of receptors occupied. When the equation is rearranged to solve for y

$$y = \frac{100Kx^n}{1 + Kx^n} \tag{3}$$

A Cartesian plot of this equation is identical in form to that shown in Figure 57-2. When y is plotted against log x instead of x, the usual sigmoid curve is obtained. Thus, it may be seen that the dose-intensity curve derives from mass action equilibrium kinetics, which in turn derive from the statistical nature of molecular interaction. The fact that dose-intensity and dose-percentage curves have the same shape shows that they involve similar statistics.

If Equation 2 is put into log form

$$\log K + n \log x = \log \frac{y}{100 - y} \tag{4}$$

a plot of log y/100 - y against log x then will yield a straight line with a slope of n; n is theoretically the number of molecules of drug that react with each molecule of receptor. At present, there are no known examples in which more than one molecule of agonist combines with a single receptor, hence, n should equal 1, universally. Nevertheless, n often deviates from 1. Deviations occur because of cooperative interactions among receptors *(cooperativity), spare receptors* (see below), amplifications in the response system *(cascades),* receptor coupling to more than one sequence (eg, to both adenylate cyclase and calcium channels), and other reasons. In these departures from $n = 1$, the slope becomes a characteristic of the mechanism of action and response system.

The probability that a molecule of drug will react with a receptor is a function of the concentration of both drug and receptor. The concentration of receptor molecules cannot be manipulated as the concentration of a drug can. But, as each molecule of drug combines with a receptor, the population of free receptors is diminished accordingly. If the drug is a competitive antagonist, it will diminish the probability of an agonist-receptor combination in direct proportion to the percentage of receptor molecules preempted by the antagonist. Consequently, the intensity of effect will be diminished. However, the probability of agonist-receptor interaction can be increased by increasing the concentration of agonist, and the intensity of effect can be restored by appropriately larger doses of agonist. Addition of more antagonist will again diminish the response, which can, again, be overcome or *surmounted* by more agonist.

Clark showed empirically and by theory that as long as the ratio of antagonist to agonist was constant, the concentration of the competitive drugs could be varied over an enormous range without changing the magnitude of the response (Fig 57-6). Since the presence of competitive antagonist only diminishes the probability of agonist-receptor combination at a given concentration of agonist and does not alter the molecularity of the reaction, it also follows that the effect of the competitive antagonist is to shift the dose-intensity curve to the right in proportion to the amount of antagonist present; neither shape nor slope of the curve is changed (Fig 57-7).

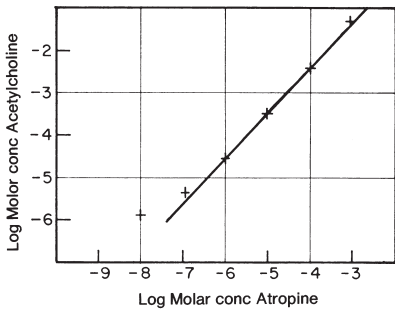


Figure 57-6. Direct proportionality of the dose of agonist (acetylcholine) to the dose of antagonist (atropine) necessary to cause a constant degree of inhibition (50%) of the response of the frog heart. (Adapted from Clark AJ. *J Physiol (London)* 1926; 61:547.)

Many refinements of the Clark formula have been made, but they will not be treated here; details and citations of relevant literature can be found in various works on receptors cited in the *Bibliography*. Several refinements are introduced to facilitate studies of competitive inhibition. The introduction of the concepts of intrinsic activity[2] and efficacy[3] required appropriate changes in mathematical treatment.

Another important concept has been added to the occupation theory, namely the concept of *spare receptors*. Clark assumed that the maximal response occurred only when the receptors were completely occupied, which does not account for the possibility that the maximum response might be limited by some step in the action-effect sequence subsequent to receptor occupation. Work with isotopically labeled agonists and antagonists and with dose-effect kinetics has shown that the maximal effect sometimes is achieved when only a small fraction of the receptors are occupied. The mathematical treatment of this phenomenon has enabled theorists to explain several puzzling observations that previously appeared to contradict occupation theory.

The classical occupation theory fails to explain several phenomena satisfactorily, and it is unable to generate a realistic model of intrinsic activity and partial agonism. A rate theory, in which the intensity of response is proportional to the rate of drug-receptor interaction instead of occupation, was proposed to explain some of the phenomena that occupation theory could not, but the rate theory was unable to provide a realistic mechanistic model of response generation, and it had other serious limitations as well.

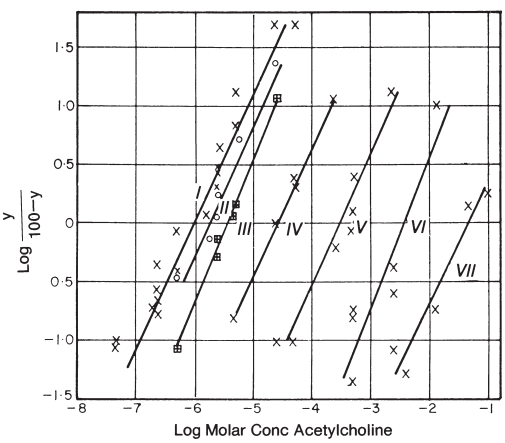


Figure 57-7. Effect of an antagonist to shift the log dose-intensity curve to the right without altering the slope. The effector is the isolated heart. *I:* no atropine; *II:* atropine, 10^{-8} M; *III:* 10^{-7} M; *IV:* 10^{-6} M; *V:* 10^{-5} M; *VI:* 10^{-4} M; *VII:* 10^{-3} M; *Y:* % of maximum intensity of response. The function log *y*/(100 - *y*) converts the log dose-intensity relationship to a straight line. (Adapted from Clark AJ. *J Physiol (London)* 1926; 61:547.)

The phenomena that neither the classical occupation nor the rate theory could explain can be explained by various theories in which the receptor can exist in at least two conformational states, one of which is the active one; the drug can react with one or more conformers. In a *two-state model*[4]

$$R \rightleftharpoons R^*$$

where R is the inactive and R* is the active conformer. The agonist combines mainly with R*, the partial agonist can combine with both R and R*, and the antagonist can combine with R, the equilibrium being shifted according to the extent of occupation of R and R*. Other variations of occupation theory treat the receptor as an aggregate of subunits that interact cooperatively.[5]

MECHANISMS OF DRUG ACTION

Drugs are distributed to many or all parts of the body by the circulation. However, they do not act everywhere; they would have extremely limited usefulness if they did. Clinically useful drugs act only on certain existing biological systems. Although drugs cannot create new systems, some drugs can temporarily or permanently damage existing functional systems that are susceptible to them, thereby producing toxic effects. Almost all drugs act more or less *selectively* on large specific proteins, glycoproteins, or lipoproteins located on the cell membrane or in the cell cytoplasm, nuclei, or other intracellular organelles. These specific proteins are referred to as *receptors*. Although they often are regarded as drug receptors, they are in reality receptors for *endogenous* substances that mediate normal biological and physiological regulatory processes.

Virtually all cells of the body have multiple receptors, since they are regulated by a variety of endogenous substances that act continuously, intermittently, or only occasionally. Similarly, cells theoretically can be influenced by a variety of drugs that act on the different receptors that the cells contain. The chemical nature of many of the endogenous substances that activate receptors is known, but new ones continue to be identified and sought. For example, the former mystery of why animals have receptors for morphine, which is produced by some species of poppy plants, was solved when endogenous opioid peptides were identified in the brain and some peripheral tissues in the mid-1970s.

Drugs that selectively activate receptors and produce the same *effects* normally produced by a respective endogenous substance are called *agonists*. Drugs that selectively block receptors are called *antagonists* because they antagonize, or block, the normal effects of the respective endogenous substance. Pure antagonists do not activate their receptors. Some experimental drugs stimulate or activate certain enzymes, but none are useful therapeutic agents because their effects are too widespread. Forskolin is one such example; it directly stimulates the enzyme adenylyl cyclase to synthesize cyclic AMP, which is a second messenger in many cellular systems throughout the body.

On the other hand, many very useful therapeutic drugs are *enzyme inhibitors*, which selectively inhibit the normal activity of only one type of enzyme, thereby reducing the ability of the enzyme to act on its normal biochemical substrate. In this context, the enzymes are the drug receptors. Although the chemical nature of receptors and enzymes and their interactions with drugs was often vague in the past, the application of new techniques in molecular biology, biochemistry, and pharmacology since the mid-1980s has made unprecedented progress in defining the structures of receptors and enzymes and the consequences of drug-receptor interactions.

TYPES OF TARGETS FOR DRUG ACTION

Drug *effects* are the result of drug *actions*. Drug action may be defined as the drug-receptor interaction, whereas drug effects are the consequences of that action. For example, the interaction of epinephrine with β-receptors in the heart sets into mo-

tion a cascade of intracellular events (actions) that lead to increases in heart rate and strength of contraction (effects). The interaction of epinephrine with α-receptors in the vasculature sets into motion a cascade of intracellular events (actions) that lead to vasoconstriction and increased blood pressure (effects).

Typical responses that involve drug-receptor interactions are those that involve agonist or antagonist interactions at a receptor. Agonists also can act through various transduction mechanisms to produce a variety of intracellular changes that alter cellular activity. Transduction mechanisms are considered in more detail near the end of this section. Agonist actions may be direct, as with acetylcholine acting on the nicotinic receptors at the neuromuscular junction to briefly open sodium channels. This produces rapid depolarization of skeletal muscle, leading to muscle contraction. Drugs also can act directly on ion channels to block their activity. For example, lidocaine *(Xylocaine)* and other local anesthetics block sodium channels in nerve fibers (axons) so that the conduction of action potentials is blocked, and the area served by those nerve fibers is anesthetized. Drugs also can act directly on ion channels to modulate their activity. The benzodiazepines, characterized by diazepam *(Valium),* produces multiple effects (sedation, hypnosis, anticonvulsant and antianxiety activity, and muscle relaxation) by *modifying* the actions of GABA on its receptors in the CNS. GABA is the predominant inhibitory neurotransmitter in the CNS, and it acts on $GABA_A$-receptor complexes by opening chloride channels on neurons to hyperpolarize them and render them less excitable. The benzodiazepines act on a different receptor on the $GABA_A$-receptor complex to enhance the actions of GABA on its receptors, thereby rendering target neurons even less excitable.

Many drugs act by inhibiting enzymes so that they cannot perform their normal functions as efficiently. One such drug, omeprazole *(Prilosec),* reduces the ability of parietal cells in the stomach to produce hydrochloric acid by inhibiting the enzyme, or proton pump, H^+, K^+-ATPase, which is found only in these parietal cells. It is used to facilitate healing of peptic ulcers and control esophageal reflux (heartburn). The body's normal enzymes also can convert false substrates into active drugs. For example, α-methyldopa *(Aldomet)* is converted into α-methylnorepinephrine by the enzymes that normally synthesize dopamine and norepinephrine from dopa. α-Methylnorepinephrine acts on brain receptors to reduce sympathetic activity to blood vessels, thereby reducing blood pressure in hypertensive patients. Antimetabolites used to treat cancer are also false substrates, which are similar in structure to endogenous metabolites involved in cell-cycle reactions but function abnormally to interfere with synthesis of essential metabolites. Some drugs are, or have been, designed to be inactive until they are converted, usually by liver drug-metabolizing enzymes such as cytochrome P450, to active drug; the inactive drug is called a *prodrug*.

Various *carriers* are used by cells to take up neurotransmitters that have been released. The actions of dopamine released from dopamine nerve terminals in the brain are terminated by reuptake into the nerve terminals by a dopamine carrier. The dopamine then is reused for neurotransmission. If the carrier is blocked by a reuptake blocker such as cocaine, dopamine concentrations between the nerve terminals and the dopamine receptors build up for a time and produce greater effects.

Finally, antibiotics and antiviral, antifungal, and antiparasitic drugs owe their selectivities to selective actions on certain biochemical processes that are essential to the offending organism but are not shared by the mammalian host. The penicillins and related antibiotics interfere with the synthesis of rigid cell walls by growing bacteria, but mammalian cells are contained only by plasma membranes and, therefore, are not affected by penicillins. Antiparasitic drugs target enzymes found only in parasites, enzymes that are indispensable only in parasites, or biochemical functions with different pharmacological properties in the parasite and the host.

RECEPTOR BINDING

Drugs that bind to certain receptors selectively at pharmacological concentrations are known as *receptor ligands;* they can be agonists or antagonists. Many drugs also bind nonselectively to nonreceptor proteins throughout the body where they exert no pharmacological actions or effects. Many drugs bind to plasma proteins, especially albumin. Albumin-bound drug can act as a reservoir for free drug, with which it is in equilibrium, and competition among drugs for plasma protein binding can lead to increased free drug levels and drug interactions as they displace one another.

Drugs and endogenous ligands or substrates bind selectively to certain receptors because of both a chemical attraction and a proper *fit* to the protein. The lock-and-key analogy provides a useful concept of proper fit. Carried a step further, an agonist fits the lock and turns it, but an antagonist only fits the lock but cannot turn it; yet, it does block entry of the agonist key. Generally, a number of drugs with both characteristics can combine with the same receptor. The study of structure-activity relationships among similar drugs and their receptors always has been an important and fruitful approach of both pharmacology and medicinal chemistry. Highly selective drugs tend to bind to only one or several closely related receptors. However, some drugs can combine with and activate or inactivate a number of different receptors that have similar structures, thereby diminishing selectivity and magnifying side effects.

The types of chemical bonds by which drugs bind to their receptors are, in decreasing order of bond strength: covalent, ionic, hydrogen, hydrophobic, and van der Waals bonds. Relatively few drugs form covalent bonds with their receptors. Covalent bonds are *irreversible* and very long-lasting; new receptors or enzymes must be synthesized to restore function, and this process takes a week or two. Most drugs rely on combinations of the other weaker bonds to bind tightly but *reversibly* to receptors. For example, the binding of acetylcholine, a relatively simple molecule, to nicotinic receptors at the neuromuscular junction, involves ionic, hydrogen, and van der Waals bonds, with ionic and hydrogen bonds being the most important. It is no accident that receptor-binding drugs are partially ionized at body pH, because receptor proteins also are partially ionized. Drugs and proteins contain positively charged nitrogen groups and negatively charged carboxyl groups that strongly attract one another and usually provide the initial drug-receptor bonds. Hydrogen bonds, formed between bound hydrogen atoms and oxygen, nitrogen, fluoride, or sulfur atoms, further orient the drug molecule to its receptor to enhance the proper fit. One or several hydrogen bonds can be involved. Hydrophobic bonds form among nonpolar ring structures (eg, benzene) or chains of methylene groups to stabilize orientation further. Finally, the very weak van der Waals forces provide some additional, electrostatic bonding over very short distances.

Drug molecules that contain asymmetrical carbon atoms can exist as stereoisomers, only one of which is oriented to bond well with its receptors. For example, the side chain of epinephrine contains an asymmetrical carbon atom in the alpha position of the side chain, with a hydroxyl group attached, permitting epinephrine to exist in D- and L- forms (mirror images). The endogenous L-form is about 1000 times more potent than the synthesized D-form because the L-form has a much greater binding affinity for its receptors because of its preferred configuration (see Chapter 28). In the past, drugs synthesized as mixtures of stereoisomers were formulated as racemic mixtures, but improved chemical separation techniques now often allow isolation of the more active isomer for formulation.

RECEPTOR STRUCTURE AND FUNCTION

The number of receptors and their subtypes continues to grow at a rapid pace as a result of identifying new endogenous ligands and applying advancing techniques to study them. De-

spite this large number, most receptors can be classified structurally and functionally into only a few basic types that are described below. No attempt is made to provide detailed descriptions of individual receptors within each category. Rather, one or two examples will suffice for each, with brief reference to some prominent types that are therapeutically relevant.

VOLTAGE-SENSITIVE CHANNELS—While not generally classified as receptors, voltage-sensitive channels contain receptors that are acted upon by drugs or toxins to block or modify their normal function. The voltage-sensitive sodium channels in axons allow initiation and conduction of action potentials in response to a voltage change in the plasma membrane. When sodium channels open, sodium ions rush into the cytoplasm, thereby causing depolarization and propagation of the action potential. The crucial component of the sodium channel is a single protein composed of a chain of about 2000 amino acids and called the α subunit. Several β subunits with minor roles are also associated with the α subunit. The α subunit has four repeating domains composed of about 250 amino acids each, and each domain contains six, α-helical, 22–to 25–amino acid, transmembrane, spanning segments. Each domain forms one of four clusters of the six membrane-spanning regions to encircle the sodium channel so formed. On end, the channel resembles 24 cylinders neatly arranged around the sodium channel that, at rest, is charged positively due to positive charges on the four transmembrane helices that surround the channel. Upon activation, these particular helices are thought to rotate upward, thereby moving the positive charges away from the channel and allowing the positive sodium ions to rush through. The channel remains open for only about 1 msec because the voltage changes attract a protein loop of the channel in the cytoplasm to shut the channel like a tether ball. Local anesthetics block the sodium channel from the cytoplasmic side by binding to receptors inside the channel. Several neurotoxins block from the outside.

Axons are repolarized by brief (~1 msec) opening of voltage-activated potassium channels that are constructed similarly to sodium channels but are composed of four identical subunits of peptide that associate in the membrane to form the potassium channel. Each subunit spans the membrane six times. It probably functions much like the sodium channel, including inactivation by a tether-ball segment of cytoplasmic peptide. Quinidine, an antiarrhythmic drug, will block this potassium channel in the heart.

Voltage-activated calcium channels of the L-type are composed of five similar protein subunits that assemble across heart muscle and vascular smooth muscle membranes to form the calcium channel. Its arrangement in the membrane is similar to that of the sodium and potassium channels. Calcium channel blockers such as verapamil *(Calan)* and nifedipine *(Procardia)* are used to treat several cardiovascular conditions by virtue of their ability to block calcium channels in the heart and blood vessels.

LIGAND-ACTIVATED ION CHANNELS—The best-characterized ligand-activated ion channel is the nicotinic receptor complex at the neuromuscular junction. As the name implies, these channels are activated by receptor ligands, in this case acetylcholine. The nicotinic receptor complex is composed of five subunit proteins with similar structures that associate across the plasma membrane to form a sodium channel. The receptor complex is formed from two α and one each of β, γ, and δ subunits (Fig 57-8). In contrast to the voltage-activated ion channels, each of the five proteins crosses the membrane only four times. The two α subunits contain the nicotinic receptors, which acetylcholine activates, and both must be activated to open the sodium channel to 6.5 Å for about 4 msec. The receptors can be blocked by neuromuscular blocking agents such as curare. The nicotinic receptors on autonomic ganglia are similar in structure but are composed of a different set of subunits, which accounts for the long-known differences in selective antagonists at the two sites.

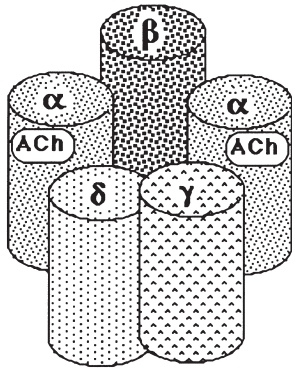


Figure 57-8. Nicotinic receptor complex.

Other ligand-activated ion channels, $GABA_A$, glycine, and glutamate, have structures that are similar to that of the nicotinic receptor complex. GABA and glycine channels are chloride channels, which permit chloride influx into neurons to produce hyperpolarization and decreased neuronal excitability. Glutamate channels are primarily sodium channels, and they also contain modifying receptors for glycine and polyamines. The $GABA_A$-receptor complex contains receptors not only for GABA but also separate receptors for benzodiazepines (eg, *Valium*), barbiturates, and steroids, which modify the actions of GABA on the chloride channel. The convulsant activity of strychnine is due solely to its ability to block glycine receptors, primarily in the brainstem and spinal cord.

G PROTEIN–COUPLED RECEPTORS—These receptors comprise a very large family of receptors that are activated by monoamines (epinephrine, norepinephrine, dopamine, and serotonin), acetylcholine (muscarinic receptors), opioids, and a host of active peptides including a number of hormones. Structurally, these receptors are single proteins, most of which are composed of chains of 350 to 550 amino acids and cross the plasma membrane seven times in a *serpentine* arrangement (Fig 57-9). Each

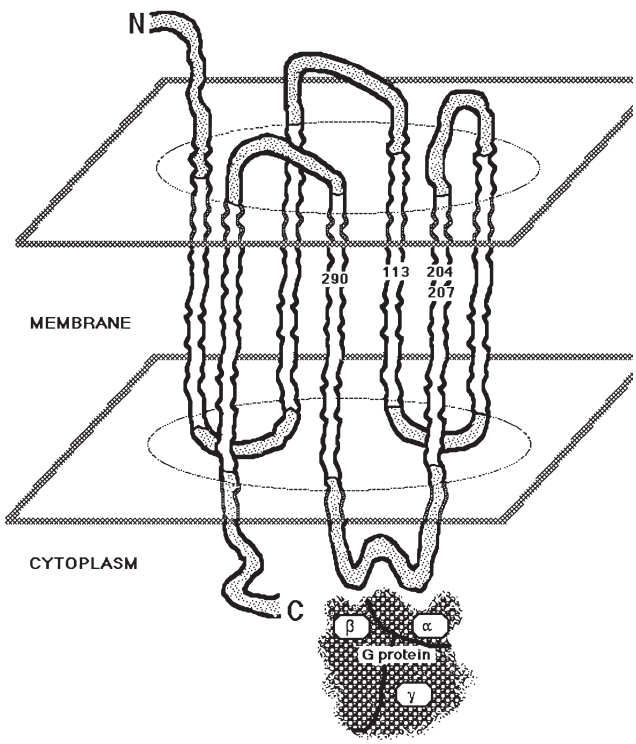


Figure 57-9. G-Protein coupled receptor complex.

of the seven transmembrane domains is composed of 22 to 30 amino acids configured into an α-helix. The third of three intracellular (cytoplasmic) loops is much longer than the other two and is responsible for coupling with the G proteins. Rather than residing at the extracellular surface of the receptor, the actual receptor-binding sites often lie *within* the membrane between the seven transmembrane domains. For example, the β-adrenergic receptor lies 11 Å below the extracellular surface, or about one-third of the distance through the membrane. The positively charged nitrogen on the side chain of the epinephrine molecule forms an ionic bond with the negatively charged carboxyl group on an aspartate amino acid (residue 113) in the third transmembrane domain (TM3). The two catechol hydroxyl groups of epinephrine form hydrogen bonds with the free hydroxyl groups of two serine amino acids at residues 204 and 207 in TM5, and the aromatic ring of epinephrine forms a hydrophobic bond with that of a phenylalanine at residue 290 in TM6. The location of G protein-coupled receptors within the membrane underscores the importance of *size and configuration* in the molecular structure of both agonists and antagonists for these receptors. Some negatively charged and peptide ligands do bind to an extracellular domain, however.

Among some families of G protein-coupled receptors there is considerable structural homology; ie, the same amino acids and the same sequences make up large portions of a number of different receptors. Consequently, a number of antagonist receptor ligands bind to these similar arrangements of amino acids in the transmembrane domains. For example, many of the antipsychotic drugs (neuroleptics) are antagonists not only at dopamine receptors, where they are thought to exert their therapeutic effects, but also at α_1-adrenergic, serotonin, histamine, and muscarinic receptors, thereby producing hypotension, sedation, blurred vision, dry mouth, and constipation as side effects.

The G proteins closely associated with the third cytoplasmic loop of the receptors are heterotrimers composed of three different subunits, α, β, and γ. Upon receptor activation, the α subunit exchanges a bound GDP for a GTP and dissociates from the $\beta\gamma$ subunits to activate a membrane enzyme such as adenylyl cyclase or to influence an ion channel. In some cases, the $\beta\gamma$ subunits may interact with the same or a different intracellular effector. The duration of action of the active GTP–α subunit is determined by the hydrolysis of GTP to GDP by a GTPase, which is intrinsic to the α subunit, and its reassociation with the $\beta\gamma$ subunits. This process is of longer duration than the association of the ligand with the ligand-G protein-coupled receptor, resulting in *amplification* of the original signal.

In the case of adenylyl cyclase activation, this enzyme synthesizes cyclic adenosine-3′,5′-monophosphate (cAMP) from ATP. As a *second messenger,* cAMP then goes on to activate one or several protein kinase As that phosphorylate one or several other proteins to produce the appropriate cellular effects. The targeted protein may be an enzyme, a transport protein, a contractile protein, or an ion channel. The specificity of these regulatory effects depends on the distinct protein substrates that are expressed in different cells (eg, liver vs smooth muscle). The actions of cAMP are terminated by several types of intracellular phosphodiesterases that convert cAMP to 5‘-AMP. Competitive inhibition of phosphodiesterases to prolong the actions of cAMP is one of the mechanisms by which caffeine produces its effects.

As if the foregoing is not sufficiently complicated, the activity of adenylyl cyclase can also be inhibited by activation of different G protein–coupled receptors. The G proteins coupled to inhibitory receptors are designated Gi proteins, as opposed to those coupled to stimulatory receptors and designated Gs proteins. Gi proteins are also heterotrimers, and receptor activation of Gi also leads to GTP binding to the α subunit and its dissociation from the $\beta\gamma$, but Gi proteins differ structurally from Gs proteins. Examples of Gs-coupled receptors are β-adrenergic, dopamine-1, histamine-2, glucagon, and ACTH. Examples of Gi-coupled receptors are α_2-adrenergic, dopamine-2, muscarinic, and opioid. A number of different Gs and Gi protein-coupled receptors can exist on the same cell, so that the activity of adenylyl cyclase can be fine-tuned between zero and maximum.

Another important group of G protein–coupled receptors activate the enzyme phospholipase C (PLC) to hydrolyze a minor component of the plasma membrane, phosphatidylinositol-4,5-biphosphate, into two second messengers, diacylglycerol (DAG) and inositol-1,4,5-triphosphate (IP3). In contrast to the cAMP systems, receptors coupled to PLC are only excitatory. Examples are α_1-adrenergic, muscarinic, Substance P, and thyrotropin-releasing hormone receptors. The second messenger DAG is confined to the membrane, where it activates a protein kinase C, of which nine distinct types have been identified. The other second messenger, IP3, diffuses through the cytosol to release calcium from intracellular stores. Calcium is involved in many cellular regulatory activities including activation of calcium-calmodulin, which regulates the activities of other enzymes including other kinases. The kinases in turn phosphorylate enzymes, ion channels, or other proteins to produce cellular effects. When the phosphoinositide and cAMP signaling systems coexist, they can oppose or complement one another in complex ways.

A third second-messenger system uses cyclic guanosine-3′,5′-monophosphate (cGMP) in intestinal mucosa and vascular smooth muscle. It is synthesized from GTP by activation of guanylyl cyclase and activates protein kinase G, which then dephosphorylates myosin light chains in vascular smooth muscle, thereby producing muscle relaxation. Agonists, eg, acetylcholine and histamine, cause the release of nitric oxide from vascular endothelial cells, which then diffuses into the smooth muscle cells to activate guanylyl cyclase. A direct receptor-mediated activation is produced by atrial natriuretic factor (ANF), a blood-borne peptide hormone. In this case, the receptor domain is outside the membrane and is connected through a single transmembrane domain to the intracellular guanylyl cyclase enzyme, which is activated by receptor binding.

TYROSINE KINASE–LINKED RECEPTORS—These receptors are composed of an extracellular receptor domain, a *single* transmembrane domain, and an intracellular catalytic domain that catalyzes phosphorylation of tyrosine residues on target proteins. Some receptors are composed of single proteins, whereas others are assembled from two subunits (eg, insulin receptors). Activation of insulin receptors triggers increased uptake of glucose and amino acids and regulates metabolism of glycogen and lipids in the cell. The catalytic actions persist for a number of minutes after insulin leaves the binding site. Several growth factors also exert their complex cellular effects by activating tyrosine kinase or similar receptors. Growth factors trigger changes in membrane transport and other metabolic events including regulation of DNA synthesis.

INTRACELLULAR RECEPTORS THAT CONTROL DNA TRANSCRIPTION—Activation of intracellular receptors for steroids (glucocorticoids, mineralocorticoids, sex steroids, vitamin D) and thyroid hormones stimulates the transcription of certain genes by binding to specific DNA sequences in the nucleus. The receptors generally are composed of a single protein with a ligand-binding domain, a DNA-binding domain, and a transcription-activating domain. In the inactivated state, the receptor protein is bound to another protein, a heat shock protein (hsp 90), which dissociates upon activation by a hormone, permitting DNA binding and transcription of mRNA, which then is translated into new protein. This process typically takes several hours, and the effects can last for days or weeks if there is a slow turnover of the newly synthesized proteins. A similar process accounts for the induction of drug-metabolizing enzymes in the liver by certain drugs and other chemicals. In this process, formation of a heterodimeric complex between a second protein and the ligand-bound receptor is required for DNA binding.

ENZYME INHIBITION—Enzymes are very large, complex proteins or associated proteins that evolved to catalyze specific

biochemical reactions that are essential to normal cellular function. A number of very selective drugs exert their effects by inhibiting particular enzymes, so that their abilities to process their normal substrates are blocked or impaired. Enzyme inhibitors can produce competitive blockade at a substrate or cofactor binding site on the enzyme. For example, the stimulant effect of digitalis glycosides on cardiac muscle contraction is mediated by competitive inhibition of a sodium pump, Na^+,K^+-ATPase, which leads indirectly to an increase in intracellular calcium to interact with contractile proteins. Other enzyme inhibitors act noncompetitively at allosteric sites (sites remote from the substrate binding site), which prevent the enzyme from performing its catalytic function. For example, aspirin binds irreversibly to a site on cyclooxygenase that is remote from the binding site for arachidonic acid, which is normally converted to prostaglandins by the enzyme. The binding of related drugs such as ibuprofen *(Advil)* is reversible. Irreversible inhibition by the formation of covalent bonds between a drug and an enzyme is typically long lasting because new enzyme must be synthesized to restore function.

RECEPTOR REGULATION—The regulation of receptor numbers or density is normally constant, as synthesis keeps pace with degradation of the proteins. However, continuous stimulation of receptors with agonists can lead to desensitization or *down-regulation* of receptor sensitivity or number. Desensitization can occur rapidly without a change in receptor number, whereas down-regulation usually implies a decline in receptor number. For example, excess use of β -adrenergic agonists for treating bronchial asthma can lead to loss of receptor sensitivity to the agonist, caused by changes in coupling mechanisms to the G proteins. Chronic blockade of receptors can lead to *up-regulation,* which, in some cases, is due to synthesis of new receptors. An example is chronic blockade of β-adrenergic receptors in the heart, in which new β -receptors are synthesized, leading to supersensitivity upon abrupt withdrawal of the blocker. Another form of supersensitivity is demonstrated by denervation of skeletal muscle, which is followed by a proliferation of nicotinic receptors within and adjacent to the neuromuscular junction.

ABSORPTION, DISTRIBUTION, AND EXCRETION

No matter by which route a drug is administered it must pass through several to many biological membranes during the processes of absorption, distribution, biotransformation, and elimination. Since membranes are traversed in all of these events, this section begins with a brief description of biological membranes and membrane processes and the relationship of the physicochemical properties of a drug molecule to penetration and transport.

STRUCTURE AND PROPERTIES OF MEMBRANES

The concept that a membrane surrounds each cell arose shortly after the cellular nature of tissue was discovered. The biological and physicochemical properties of cells seemed in accord with this view. Microchemical, x-ray diffraction, electron microscopic, nuclear magnetic resonance, electron spin resonance, and other investigations have established the nature of the plasma, mitochondrial, nuclear, and other cell membranes. The description of the plasma membrane that follows is much oversimplified, but it will suffice to provide a background for an understanding of drug penetration into and through membranes.

STRUCTURE AND COMPOSITION—The cell membrane has been described as a bimolecular layer of lipid material entrained between two parallel monomolecular layers of protein. However, the protein does not make continuous layers, but rather is sporadically scattered over the surfaces, like icebergs; ie, much of the protein is below the surface. In Figure 57-10 the lipid layers are represented as a somewhat orderly, closely packed, lamellar array of phospholipid molecules associated tail-to-tail, each *tail* being an alkyl chain or steroid group, and the *heads* being polar groups, including the glycerate moieties, with their polar ether and carbonyl oxygens and phosphate with attached polar groups. In reality, the lamellar portion is probably not so orderly, since its composition is quite complex. Chains of fatty acids of different degrees of saturation and cholesterol cannot array themselves in simple parallel arrangements. Furthermore, the polar heads will assume a number of orientations depending upon the substances and groups involved. Moreover, the lamellar portion is penetrated by large globular proteins, the interior of which, like the lipid layers, has a high hydrophobicity, and some fibrous proteins.

The plasma membrane appears to be asymmetrical. The lipid composition varies from cell type to cell type and perhaps from site to site on the same membrane. There are, for example, differences between the membrane of the endoplasmic reticulum and the plasma membrane, even though the membranes are coextensive. Where membranes are double, the inner and outer layers may differ considerably; the inner and outer membranes of mitochondria have been shown to have strikingly different compositions and properties. Some authorities have expressed doubt as to the existence of the protein layers in biological membranes, although the evidence is preponderantly in favor of at least an outer glycoprotein coat. Sugar moieties also are attached to the outer proteins, most often to the asparagine residue. These sugar moieties are important to cellular and immunological recognition and adhesion and have other functions as well.

The cell membrane appears to be perforated by water-filled pores of various sizes, varying from about 4 to 10 Å, most of which are about 7 Å. Probably all major ion channels are through the large globular proteins that traverse the membrane. Through these pores pass inorganic ions and small organic molecules. Since sodium ions are more hydrated than potassium and chloride ions, they are larger and do not pass as freely through the pores as potassium and chloride. The vascular endothelium appears to have pores at least as large as 40 Å, but these seem to be interstitial passages rather than transmembrane pores. Lipid molecules small enough to pass through the pores may do so, but

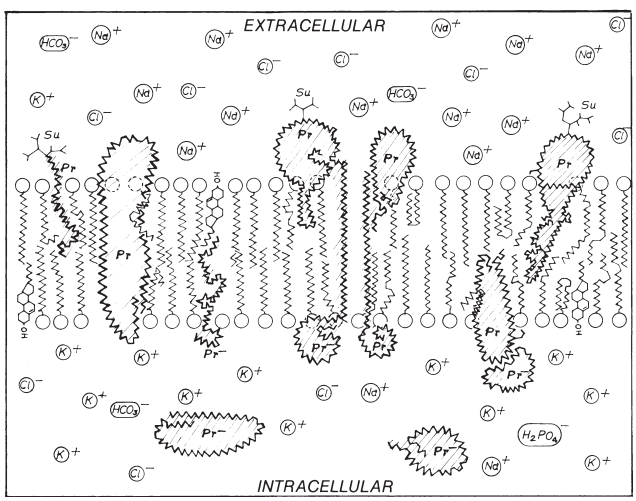


Figure 57-10. Simplified cross section of a cell membrane (components are not to scale). The lipid interior of the lamellar portion of the membrane consists of various phospholipids, fatty acids, cholesterol, and other steroids. Ions are indicated to illustrate differences in size relative to the channel. *Pr,* protein; *Su,* sugar.

they have a higher probability of entering into the lipid layer, from where they will equilibrate chemically with the interior of the cell. From work on monolayers, some researchers contend that it is not necessary to postulate pores to explain the permeability to water and small water-soluble molecules.

STRATUM CORNEUM—Although the stratum corneum is not a membrane in the same sense as a cell membrane, it offers a barrier to diffusion, which is of significance in the topical application of drugs. The stratum corneum consists of several layers of dead, keratinized, cutaneous epithelial cells enmeshed in a matrix of keratin fibers and bound together with cementing desmosomes and penetrating tonofibrils of keratin. Varying amounts of lipids and fatty acids from dying cells, sebum, and sweat are contained among the dead squamous cells. Immediately beneath the layer of dead cells and above the viable epidermal epithelial cells is a layer of keratohyaline granules and various water-soluble substances, such as α-amino acids, purines, monosaccharides, and urea.

Both the upper and lower layers of the stratum corneum are involved in the cutaneous barrier to penetration. The barrier to penetration from the surface is in the upper layers for water-soluble substances and the lower layers for lipid-soluble substances, and the barrier to the outward movement of water is in the lowest layer.

MEMBRANE POTENTIALS—Across the cell membrane there exists an electrical potential, always negative on the inside and positive on the outside. If a cell did not have special-membrane electrolyte-transport processes, its membrane potential would be mainly the result of the Donnan equilibrium (see Chapter 14) consequent to the semipermeability of the membrane. Such potentials generally lie between 2 and 5 mV.

A cell with a membrane across which diffusible electrolyte distribution is purely passive would be expected to have a high internal concentration of sodium, which is true for the erythrocytes of some species. However, the interior of most cells is high in potassium and low in sodium, as depicted in Figure 57-10. This unequal distribution of cations attests to special electrolyte-transport processes and to differential permeabilities of diffusible ions, so that the membrane potential is higher than that which would result from a purely passive Donnan distribution. In nerve tissue or skeletal and cardiac muscle, the membrane potential ranges upward to about 90 mV. The electrical gradient is on the order of 50,000 V/cm, because of the extreme thinness of the membrane. Obviously, such an intense potential gradient will influence strongly the transmembrane passages of charged drug molecules.

DIFFUSION AND TRANSPORT

Transport is the movement of a drug from one place to another within the body. The drug may diffuse freely in uncombined form with a kinetic energy appropriate to its thermal environment, or it may move in combination with extracellular or cellular constituents, sometimes in connection with energy-yielding processes that allow the molecule or complex to overcome barriers to simple diffusion.

SIMPLE NONIONIC DIFFUSION AND PASSIVE TRANSPORT—Molecules in solution move in a purely random fashion, provided they are not charged and moving in an electrical gradient. Such random movement is called *diffusion;* if the molecule is uncharged, it is called *nonionic diffusion*.

In a population of drug molecules, the probability that during unit time any drug molecule will move across a boundary is directly proportional to the number of molecules adjoining that boundary and, therefore, to the drug concentration. Except at dilutions so extreme that only a few molecules are present, the actual rate of movement (molecules/unit time) is directly proportional to the probability and, therefore, to the concentration. Once molecules have passed through the boundary to the opposite side, their random motion may cause some to return and others to continue to move further away from the boundary. The rate of return is likewise proportional to the concentration on the opposite side of the boundary. It follows that although molecules are moving in both directions, there will be a net movement from the region of higher to that of lower concentration, and the net transfer will be proportional to the concentration differential. If the boundary is a membrane, which has both substance and dimension, the rate of movement is also directly proportional to the permeability and inversely proportional to the thickness. These factors combine into Fick's law of diffusion,

$$\frac{dQ}{dt} = \frac{\overline{D}A(C_1 - C_2)}{x} \tag{5}$$

where Q is the net quantity of drug transferred across the membrane, t is time, C_1 is the concentration on one side and C_2 that on the other, x is the thickness of the membrane, A is the area, and $\overline{D}$ is the diffusion coefficient, related to permeability. Since a biological membrane is heterogeneous, with pores of different sizes and probably with varying thickness and composition, both $\overline{D}$ and x probably vary from place to place. Nevertheless, some mean values can be assumed.

It is customary to combine the membrane factors into a single constant, called a permeability constant or coefficient, P, so that $P = \overline{D}/x$, and A in Equation 5 has unit value. The rate of net transport (diffusion) across the membrane then becomes

$$\frac{dQ}{dt} = P(C_1 - C_2) \tag{6}$$

As diffusion continues, C_1 approaches C_2, and the net rate, dQ/dt, approaches zero in exponential fashion, characteristic of a first-order process. Equilibrium is defined as that state in which $C_1 = C_2$. The equilibrium is, of course, dynamic, with equal numbers of molecules being transported in each direction during unit time. If water also is moving through the membrane, it may either facilitate the movement of drug or impede it, according to the relative directions of movement of water and drug; this effect of water movement is called *solvent drag*.

IONIC OR ELECTROCHEMICAL DIFFUSION—If a drug is ionized, the transport properties are modified. The probability of penetrating the membrane is still a function of concentration, but it is also a function of the potential difference or electrical gradient across the membrane. A cationic drug molecule will be repelled from the positive charge on the outside of the membrane, and only those molecules with a high kinetic energy will pass through the ion barrier. If the cation is polyvalent, it may not penetrate at all.

Once inside the membrane, a cation simultaneously will be attracted to the negative charge on the intracellular surface of the membrane and repelled by the outer surface; it is said to be moving along the *electrical gradient*. If it also is moving from a higher toward a lower concentration, it is said to be moving along its *electrochemical gradient,* which is the sum of the influences of the electrical field and the concentration differential across the membrane.

Once inside the cell, cations will tend to be kept inside by the attractive negative charge on the interior of the cell, and the intracellular concentration of drug will increase until, by sheer numbers of accumulated drug particles, the outward diffusion or mass escape rate equals the inward transport rate, and electrochemical equilibrium is said to have occurred. At electrochemical equilibrium at body temperature (37°), ionized drug molecules will be distributed according to the Nernst equation,

$$\pm \log \frac{C_0}{C_i} = \frac{ZE}{61} \tag{7}$$

where C_o is the molar extracellular, and C_i the intracellular, concentration; Z is the number of charges per molecule, and E is the membrane potential in millivolts. Log C_o/C_i is positive when the molecule is negatively charged and negative when the molecule is positively charged.

FACILITATED DIFFUSION—Sometimes a substance moves more rapidly through a biological membrane than can be accounted for by the process of simple diffusion. This acceler-

ated movement is termed *facilitated diffusion*. It is thought to be due to the presence of a special molecule within the membrane, called a *carrier,* with which the transported substance combines. There is considered to be greater permeability to the carrier-drug complex than to the drug alone, so that the transport rate is enhanced. After the complex traverses the membrane, it dissociates. The carrier must either return to the original side of the membrane to be reused or constantly be produced on one side and eliminated on the other for the carrier process to be continuous. Many characteristics of facilitated diffusion, formerly attributed to ion carriers, can be explained by ion exchange. Although facilitated diffusion resembles active transport, below, in its dependence upon a continuous source of energy, it differs in that facilitated diffusion will only transport a molecule along its electrochemical gradient.

ACTIVE TRANSPORT—Active transport may be defined as energy-dependent movement of a substance through a biological membrane against an electrochemical gradient. It is characterized by

1. The substance is transported from a region of lower to one of higher electrochemical activity.
2. Metabolic poisons interfere with transport.
3. The transport rate approaches an asymptote (ie, saturates) as concentration increases.
4. The transport system usually shows a requirement for specific chemical structures.
5. Closely related chemicals are competitive for the transport system.

Many drugs are secreted from the renal tubules into urine, from liver cells into bile or blood, from intestinal cells into the lumen of the GI tract, or from the cerebrospinal fluid into blood by active transport, but the role of active transport of drugs in the distribution into most body compartments and tissues is less well documented. Active transport is required for the penetration of a number of sympathomimetics into neural tissue and for the movement of several anticancer drugs across cell membranes.

PINOCYTOSIS AND EXOCYTOSIS—Many, perhaps all, cells are capable of a type of phagocytosis called *pinocytosis*. The cell membrane has been observed to invaginate into a saccular structure containing extracellular materials and then pinch off the saccule at the membrane, so that the saccule remains as a vesicle or vacuole within the interior of the cell. Since metabolic activity is required and since an extracellular substance may be transported against an electrochemical gradient, pinocytosis shows some of the same characteristics as active transport. However, pinocytosis is relatively slow and inefficient compared with most active transport, except in GI absorption, in which pinocytosis can be of considerable importance.

It is not known to what extent pinocytosis contributes to the transport of most drugs, but many macromolecules and even larger particles can be absorbed by the gut. Pinocytosis probably explains the oral efficacy of the Sabin polio vaccine. Some drugs themselves affect pinocytosis; eg, adrenal glucocorticoids markedly inhibit the process in macrophages and other cells involved in inflammation.

Exocytosis is more or less the reverse of pinocytosis. Granules, vacuoles, or other organelles within the cell move to the cell membrane, fuse with it, and extrude their contents into the interstitial space.

PHYSICOCHEMICAL FACTORS IN PENETRATION

Drugs and other substances may traverse the membrane primarily either through the pores or by dissociation into the membrane lipids and subsequent diffusion from the membrane into the cytosol or other fluid on the far side of the membrane. The physicochemical prerequisites differ according to which route is taken. To pass through the pores, the *diameter* of the molecule must be smaller than the pore, but the molecule can be longer than the pore diameter. The probability that a long, thin molecule will be oriented properly is low unless there is also bulk flow, and the transmembrane passage of large molecules is slow.

Water-soluble molecules with low lipid solubility usually are thought to pass through the membrane mainly via the pores and, to a small extent, by pinocytosis, although work with lipid monolayers suggests that small, water-soluble molecules also may be able to pass readily through the lipid, and the necessity of postulating the existence of pores has been questioned. Nevertheless, experimental data on penetration overwhelmingly favor the concept of passage of water-soluble, lipid-insoluble substances through pores. If there is a membrane carrier or active transport system, a low solubility of the drug in membrane lipids is no impediment to penetration, since the drug-carrier complex is assumed to have an appropriate solubility, and energy from an active transport system enables the drug to penetrate the energy barrier *imposed by the lipids*. Actually, the lipids are not an important energy barrier; rather, the barrier is the force of attraction of the solvent water for its dipolar-to-polar solute, so that it is difficult for the solute to leave the water and enter the lipid.

Drugs with a high solubility in the membrane lipids pass easily through the membrane. Even when their dimensions are small enough to permit passage through pores, lipid-soluble drugs primarily pass through the membrane lipids, not only because chemical partition favors the lipid phase but also because the surface area occupied by pores is only a small fraction of the total membrane area.

LIPID SOLUBILITY AND PARTITION COEFFICIENTS—As early as 1902, Overton investigated the importance of lipid solubility to the penetration and absorption of drugs. Eventually, it was recognized that more important than lipid solubility was the lipid-water distribution coefficient; ie, a high lipid solubility does not favor penetration unless the water solubility is low enough so that the drug is not entrained in the aqueous phase.

In Figure 57-11 is illustrated the relationship between the chloroform-water partition coefficient and the colonic absorption of barbiturates. Chloroform probably is not the optimal lipid solvent for such a study, and natural lipids from nerve or other tissues have been shown to be superior in the few instances in which they have been employed. Nevertheless, the correlation shown in the figure is a convincing one.

When the water solubility of a substance is so low that a significant concentration in water or extracellular fluid cannot be achieved, absorption may be negligible in spite of a favorable partition coefficient. Hence, mineral oil, petrolatum, etc, virtually are unabsorbed. The optimal partition coefficient for permeation of the skin appears to be lower than that for the permeation of the cell membrane, perhaps being as low as one.

DIPOLARITY, POLARITY, AND NONIONIC DIFFUSION—The partition coefficient of a drug depends upon the polarity and the size of the molecule. Drugs with a high dipole moment, even though un-ionized, have a low lipid solubility and, hence, penetrate poorly. An example of a highly dipolar substance with a low partition coefficient, which does not penetrate into cells, is sulfisoxazole. Sulfadiazine is somewhat less dipolar, has a chloroform-water partition coefficient 10 times that of sulfisoxazole, and readily penetrates cells. Ionization not only diminishes lipid solubility greatly but also may impede passage through charged membranes (see *Ionic Diffusion*).

It often is stated that ionized molecules do not penetrate membranes, except for ions of small diameter. This is not necessarily true, because of the presence of membrane carriers for some ions, which effectively may shield or neutralize the charge (ion-pair formation). The renal tubular transport systems, which transport such obligate ions as tetraethylammonium, probably form ion-pairs. Furthermore, if an ionized molecule has a large nonpolar moiety such that an appreciable lipid solubility is imparted to the molecule in spite of the charge, the drug may penetrate, although usually at a slow rate. For example, various morphinan derivatives are absorbed passively from the stomach even though they are ionized completely at

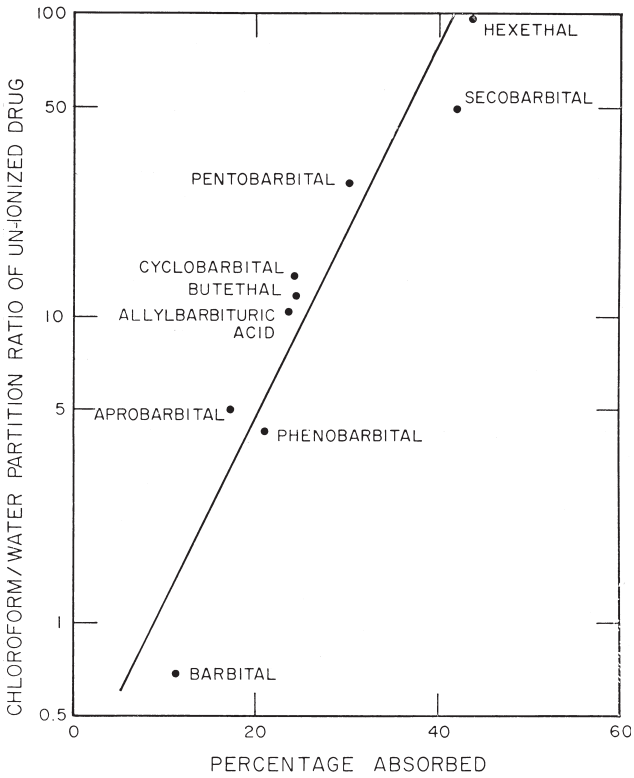


Figure 57-11. The relationship of absorption of the un-ionized forms of drugs from the colon of the rat to the chloroform:water partition coefficient. (From Schanker LS. *Adv Drug Res* 1964; 1:71.)

the pH of gastric fluid. Nevertheless, when a drug is a weak acid or base, the un-ionized form, with a favorable partition coefficient, passes through a biological membrane so much more readily than the ionized form that for all practical purposes, only the un-ionized form is said to pass through the membrane. This has become known as the *principle of nonionic diffusion*.

This principle is the reason that only the concentrations of the un-ionized form of the barbiturates are plotted in Figure 57-11.

For the purpose of further illustrating the principle, Table 57-1 is provided.[7] In the table, the permeability constants for penetration into the cerebral spinal fluid of rats are higher for un-ionized drugs than for ionized ones. The apparent exceptions—barbital, sulfaguanidine, and acetylaminoantipyrine—may be explained by the dipolarity of the un-ionized molecules. With barbital, the two lipophilic ethyl groups are too small to compensate for the considerable dipolarity of the un-ionized barbituric acid ring; also it may be seen that barbital is appreciably ionized, which contributes to the relatively small permeability constant. Sulfaguanidine and acetylaminoantipyrine are both very polar molecules. Mecamylamine also might be considered an exception, since it shows a modest permeability even though strongly ionized; there is no dipolarity in mecamylamine except in the amino group.

Absorption of Drugs

Absorption is the process of movement of a drug from the site of application into the extracellular compartment of the body. Inasmuch as there is a great similarity among the various membranes that a drug may pass through to gain access to the extracellular fluid, it might be expected that the particular site of application (or *route*) would make little difference to the successful absorption of the drug. In actual fact, it makes a great deal of difference; many factors, other than the structure and composition of the membrane, determine the ease with which a drug is absorbed. These factors are discussed in the following sections, along with an account of the ways that drug formulations may be manipulated to alter the ability of a drug to be absorbed readily.

ROUTES OF ADMINISTRATION

Drugs may be administered by many different routes. The various routes include oral, rectal, sublingual or buccal, parenteral, inhalation, and topical. The choice of a route depends upon both convenience and necessity.

ORAL ROUTE—This is obviously the most convenient route for access to the systemic circulation, providing that various factors do not militate against this route. Oral administration does not always give rise to sufficiently high plasma concentrations to be effective; some drugs are absorbed unpredictably or erratically; patients occasionally have an absorption malfunction. Drugs may not be given by mouth to patients with GI intolerance or who are in preparation for anesthesia or who have had GI surgery. Oral administration also is precluded in coma.

RECTAL ROUTE—Drugs that ordinarily are administered by the oral route usually can be administered by injection

Table 57-1. Rates of Entry of Drugs in CSF and the Degrees of Ionization of Drugs at pH 7.4[7]

DRUG/CHEMICAL	% BINDING TO PLASMA PROTEIN	PK_a[a]	% UN-IONIZED AT pH 7.4	PERMEABILITY CONSTANT ($P\ min^{-1}$) ± S.E.
Drugs mainly ionized at pH 7.4				
5-Sulfosalicylic acid	22	(strong)	0	<0.0001
N-Methylnicotinamide	<10	(strong)	0	0.0005 ± 0.00006
5-Nitrosalicylic acid	42	2.3	0.001	0.001 ± 0.0001
Salicylic acid	40	3.0	0.004	0.006 ± 0.0004
Mecamylamine	20	11.2	0.016	0.021 ± 0.0016
Quinine	76	8.4	9.09	0.078 ± 0.0061
Drugs mainly un-ionized at pH 7.4				
Barbital	<2	7.5	55.7	0.026 ± 0.0022
Thiopental	75	7.6	61.3	0.50 ± 0.051
Pentobarbital	40	8.1	83.4	0.17 ± 0.014
Aminopyrine	20	5.0	99.6	0.25 ± 0.020
Aniline	15	4.6	99.8	0.40 ± 0.042
Sulfaguanidine	6	>10.0[b]	>99.8	0.003 ± 0.0002
Antipyrine	8	1.4	>99.9	0.12 ± 0.013
N-Acetyl-4-aminoantipyrine	<3	0.5	>99.9	0.012 ± 0.0010

[a] The dissociation constant of both acids and bases is expressed as the pK_a, the negative logarithm of the acidic dissociation constant.
[b] Sulfaguanidine has a very weakly acidic group ($pK_a > 10$) and two very weakly basic groups (pK_a 2.75 and 0.5). Consequently, the compound is almost completely undissociated at pH 7.4.

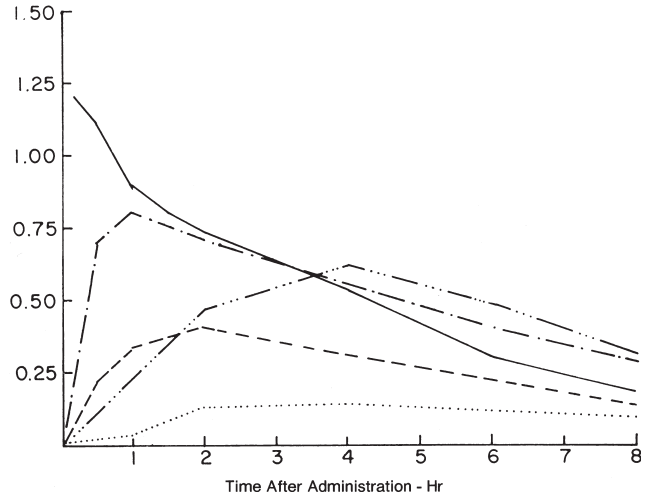


Figure 57-12. Blood concentration in mg/100 mL of theophylline (ordinate) following administration to humans of aminophylline in the amounts and by the routes indicated. Doses: per 70 kg. Theophylline-ethylenediamine by various routes:——intravenous, 0.5 g;–·–·–retention enema, 0.5 g;—• • •—• • •—oral tablets-Pl, 0.5 g; - - - oral tablets-Pl, 0.3 g; • • • • rectal suppository, 0.5 g. (Adapted Truitt EB, et al. *J Pharmacol Exp Ther* 1950; 100:309.)

or by the alternative *lower enteral* route, through the anal portal into the rectum or lower intestine. With regard to the latter, *rectal suppositories* or *retention enemas* formerly were used quite frequently, but their popularity has abated somewhat, owing to improvements in parenteral preparations. Nevertheless, they continue to be valid and, sometimes, very important ways of administering a drug, especially in pediatrics and geriatrics. In Figure 57-12[8] the availability of a drug by retention enema may be compared with that by the intravenous and oral routes and rectal suppository administration. It is apparent that the retention enema may be a very satisfactory means of administration but that rectal suppositories may be inadequate when rapid absorption and high plasma levels are required. The illustration is not intended to lead the reader to the conclusion that a retention enema always will give more prompt and higher blood levels than the oral route, for converse findings for the same drug have been reported,[9] but rather to show that the retention enema may offer a useful substitute for the oral route.

SUBLINGUAL OR BUCCAL ROUTE—Even though an adequate plasma concentration eventually may be achievable by the oral route, it may rise much too slowly for use in some situations when a rapid response is desired. In such situations parenteral therapy usually is indicated. However, the patients with angina pectoris may get quite prompt relief from an acute attack by the *sublingual* or *buccal* administration of nitroglycerin, so that parenteral administration may be avoided. When only small amounts of drugs are required to gain access to the blood, the buccal route may be very satisfactory, providing the physicochemical prerequisites for absorption by this route are present in the drug and dosage form. Only a few drugs may be given successfully by this route.

PARENTERAL ROUTES—These routes, by definition, include any route other than the oral-GI (enteral) tract, but in common medical usage the term excludes topical administration and includes only various hypodermic routes. Parenteral administration includes the intravenous, intramuscular, and subcutaneous routes. Parenteral routes may be employed whenever enteral routes are contraindicated (see above) or inadequate.

The *intravenous* route may be preferred on occasion, even when a drug may be well absorbed by the oral route. There is no delay imposed by absorption before the administered drug reaches the circulation, and blood levels rise virtually as rapidly as the time necessary to empty the syringe or infusion bottle. Consequently, the intravenous route is the preferred route when an emergency calls for an immediate response.

In addition to the rapid rise in plasma concentration of drug, another advantage of intravenous administration is the greater predictability of the peak plasma concentration, which, with some drugs, can be calculated with a fair degree of precision. Smaller doses generally are required by the intravenous than by other routes, but this usually affords no advantage, inasmuch as the sterile injectable dosage form costs more than enteric preparations, and the requirements for medical or paramedical supervision of administration also may add to the cost and inconvenience.

Because of the rapidity with which drug enters the circulation, dangerous side effects to the drug may occur, which are often not extant by other routes. The principal untoward effect is a depression of cardiovascular function, which is often called *drug shock*. Consequently, some drugs must be given quite slowly to avoid vasculotoxic concentrations of drug in the plasma. Acute, serious, allergic responses also are more likely to occur by the intravenous route than by other routes.

Many drugs are too irritant to be given by the oral, intramuscular, or subcutaneous route and must, of necessity, be given intravenously. However, such drugs also may cause damage to the veins (phlebitis) or, if extravasated, cause necrosis (slough) around the injection site. Consequently, such irritant drugs may be diluted in isotonic solutions of saline, dextrose, or other media and given by slow infusion, providing that the slower rate of delivery does not negate the purpose of the administration in emergency situations.

Absorption by the *intramuscular route* is relatively fast, and this parenteral route may be used when an immediate effect is not required but a prompt effect is desirable. Intramuscular deposition also may be made of certain repository preparations, rapid absorption not being desired. Absorption from an intramuscular depot is more predictable and uniform than from a subcutaneous site.

Irritation around the injection site is a frequent accompaniment of intramuscular injection, depending upon the drug and other ingredients. Because of the dangers of accidental intravenous injection, medical supervision generally is required. Sterilization is necessary.

In *subcutaneous* administration the drug is injected into the connective tissue just below the skin. Absorption is slower than by the intramuscular route but, nevertheless, may be prompt with many drugs. Often, however, absorption by this route may be no faster than by the oral route. Therefore, when a fairly prompt response is desired with some drugs, the subcutaneous route may not offer much advantage over the oral route, unless for some reason the drug cannot be given orally.

The slower rate of absorption by the subcutaneous route is usually the reason why the route is chosen, and the drugs given by this route are usually those in which it is desired to spread the action out over a number of hours, to avoid either too intense a response, too short a response, or frequent injections. Examples of drugs given by this route are insulin and sodium heparin, neither of which is absorbed orally, and both of which should be absorbed slowly over many hours. In the treatment of asthma, epinephrine usually is given subcutaneously to avoid the dangers of rapid absorption and consequent dangerous cardiovascular effects. Many repository preparations, including tablets or pellets, are given subcutaneously. As with other parenteral routes, irritation may occur. Sterile preparations also are required. However, medical supervision is not required always and self-administration by this route is customary with certain drugs, such as insulin.

Intradermal injection, in which the drug is injected into, rather than below, the dermis, is rarely employed, except in certain diagnostic and test procedures, such as screening for allergic or local irritant responses.

Occasionally, even by the intravenous route, it is not possible, practical, or safe to achieve plasma concentrations high enough so that an adequate amount of drug penetrates into special compartments, such as the cerebrospinal fluid, or various cavities, such as the pleural cavity. The brain is especially difficult to penetrate with water-soluble drugs. The name *blood-brain barrier* is applied to the impediment to penetration. When drugs do penetrate, the choroid plexus often secretes them back into the blood very rapidly, so that adequate levels of drugs in the cerebrospinal fluid may be difficult to achieve. Consequently, *intrathecal* or *intraventricular* administration may be indicated.

Body cavities such as the pleural cavity normally are wetted by a small amount of effusate that is in diffusion equilibrium with the blood and, hence, is accessible to drugs. However, infections and inflammations may cause the cavity to fill with serofibrinous exudate that is too large to be in rapid diffusion equilibrium with the blood. *Intracavitary* administration, thus, may be required. It is extremely important that sterile, nonirritating preparations be used for intrathecal or intracavitary administration.

INHALATION ROUTE—Inhalation may be employed for delivering gaseous or volatile substances into the systemic circulation, as with most general anesthetics. Absorption is virtually as rapid as the drug can be delivered into the alveoli of the lungs, since the alveolar and vascular epithelial membranes are quite permeable, blood flow is abundant, and there is a very large surface for absorption.

Aerosols of nonvolatile substances also may be administered by inhalation, but the route is used infrequently for delivery into the systemic circulation because of various factors that contribute to erratic or difficult-to-achieve blood levels. Whether or not an aerosol reaches and is retained in pulmonary alveoli depends critically upon particle size. Particles larger than 1 μm in diameter tend to settle in the bronchioles and bronchi, whereas particles smaller than 0.5 μm fail to settle and mainly are exhaled. Aerosols are employed mostly when the purpose of administration is an action of the drug upon the respiratory tract itself. An example of a drug commonly given as an aerosol is isoproterenol, which is employed to relax the bronchioles during an asthma attack.

TOPICAL ROUTE—Topical administration is employed to deliver a drug at, or immediately beneath, the point of application. Although occasionally enough drug is absorbed into the systemic circulation to cause systemic effects, absorption is too erratic for the topical route to be used routinely for systemic therapy. However, various transdermal preparations of nitroglycerin and clonidine are employed quite successfully for systemic use. Some investigations with aprotic solvent vehicles such as dimethyl sulfoxide (DMSO) also have generated interest in topical administration for systemic effects. A large number of topical medicaments are applied to the skin, although topical drugs are also applied to the eye, nose, throat, ear, vagina, etc.

In man, percutaneous absorption probably occurs mainly from the surface. Absorption through the hair follicles occurs, but the follicles in man occupy too small a portion of the total integument to be of primary importance. Absorption through sweat and sebaceous glands generally appears to be minor. When the medicament is rubbed on vigorously, the amount of the preparation that is forced into the hair follicles and glands is increased. Rubbing also forces some material through the stratum corneum without molecular dispersion and diffusion through the barrier. Rather large particles of substances such as sulfur have been demonstrated to pass intact through the stratum corneum. When the skin is diseased or abraded, the cutaneous barrier may be disrupted or defective, so that percutaneous absorption may be increased. Since much of a drug that is absorbed through the epidermis diffuses into the circulation without reaching a high concentration in some portions of the dermis, systemic administration may be preferred in lieu of, or in addition to, topical administration.

FACTORS THAT AFFECT ABSORPTION

In addition to the physicochemical properties of drug molecules and biological membranes, various factors affect the rate of absorption and determine, in part, the choice of route of administration.

CONCENTRATION—It is self-evident that the concentration, or, more exactly, the thermodynamic activity, of a drug in a drug preparation will have an important bearing upon the rate of absorption, since the rate of diffusion of a drug away from the site of administration is directly proportional to the concentration. Thus, a 2% solution of lidocaine will induce local anesthesia more rapidly than a 0.2% solution. However, drugs administered in solid form are not absorbed necessarily at the maximal rate (see *Physical State of Formulation and Dissolution Rate,* below).

After oral administration the concentration of drugs in the gut is a function of the dose, but the relationship is not necessarily linear. Drugs with a low aqueous solubility (eg, digitoxin) quickly saturate the GI fluids, so that the rate of absorption tends to reach a limit as the dose is increased. The peptizing and solubilizing effects of bile and other constituents of the GI contents assist in increasing the rate of absorption but are in themselves somewhat erratic. Furthermore, many drugs affect the rates of gastric, biliary, and small intestinal secretion, which causes further deviations from a linear relationship between concentration and dose.

Drugs that are administered subcutaneously or intramuscularly also may not always show a direct linear relationship between the rate of absorption and the concentration of drug in the applied solution, because osmotic effects may cause dilution or concentration of the drug, if the movement of water or electrolytes is different from that of the drug. Whenever possible, drugs for hypodermic injection are prepared as isotonic solutions. Some drugs affect the local blood flow and capillary permeability, so that at the site of injection there may be a complex relationship of concentration achieved to the concentration administered.

PHYSICAL STATE OF FORMULATION AND DISSOLUTION RATE—The rate of absorption of a drug may be affected greatly by the rate at which the drug is made available to the biological fluid at the site of administration. The intrinsic physicochemical properties, such as solubility and the thermodynamics of dissolution, are only some of the factors that affect the rate of dissolution of a drug from a solid form. Other factors include not only the unavoidable interactions among the various ingredients in a given formulation but also deliberate interventions to facilitate dispersion (eg, comminution, Chapter 38 and dissolution, Chapter 35) or retard it (eg, coatings, Chapter 46 and slow-release formulations, Chapter 47). There also are factors that affect the rate of delivery from liquid forms. For example, a drug in a highly viscous vehicle is absorbed more slowly from the vehicle than a drug in a vehicle of low viscosity; in oil-in-water emulsions the rate depends upon the partition coefficient. These manipulations are the subject of biopharmaceutics (see Chapter 47).

AREA OF ABSORBING SURFACE—The area of absorbing surface is an important determinant of the rate of absorption. To the extent that the therapist must work with the absorbing surfaces available in the body, the absorbing surface is not subject to manipulation. However, the extent to which the existing surfaces may be used is subject to variation. In those rare instances in which percutaneous absorption is intended for systemic administration, the entire skin surface is available.

Subsequent to subcutaneous or intramuscular injections, the site of application may be massaged to spread the injected fluid from a compact mass to a well-dispersed deposit. Alternatively, the dose may be divided into multiple small injections, although this recourse is generally undesirable.

The different areas for absorption afforded by the various routes account, in part, for differences in the rates of absorption by those routes. The large alveolar surface of the lungs allows

extremely rapid absorption of gases, vapors, and properly aerosolized solutions; with some drugs the rate of absorption may be nearly as fast as with intravenous injection. In the gut the small intestine is the site of the fastest, and hence most, absorption because of the small lumen and highly developed villi and microvilli; the stomach has a relatively small surface area, so that even most weak acids are absorbed predominately in the small intestine despite a pH partition factor that should favor absorption from the stomach (see *The pH Partition Principle*).

VASCULARITY AND BLOOD FLOW—Although the thermal velocity of a freely diffusible, average drug molecule is on the order of meters per second, in solution the rate at which it will diffuse away from a reference point will be much slower. Collisions with water and/or other molecules that cause a random motion, and the forces of attraction between the drug and water or other molecules, slow the net mean velocity.

The time taken to traverse a given distance is a function of the square of the distance; on average it would take about 0.01 sec for a net outward movement of 1 μm, 1 sec for 10 μm, 100 sec for 100 μm, etc. In a highly vascular tissue, such as skeletal muscle, in which there may be more than 1000 capillaries/mm^2 of cross-section, a drug molecule would not have to travel more than a few microns, hence less than a second on average, to reach a capillary from a point of extravascular injection.

Once the drug reaches the blood, diffusion is not important to transport and the rate of blood flow determines the movement. The velocity of blood flow in a capillary is about 1 mm/sec, which is 100 times faster than the mean net velocity of drug molecules 1 mm away from their injection site. The velocity of blood flow is even faster in the larger vessels. Overall, less than a minute is required to distribute drug molecules from the capillaries at the injection site to the rest of the body.

From the above discussion it follows that absorption is most rapid in the vascular tissues. Drugs are absorbed more rapidly from intramuscular sites than from less vascular subcutaneous sites, etc. Despite the small absorbing surface for buccal or sublingual absorption, the high vascularity of the buccal, gingival, and sublingual surfaces favors an unexpectedly high rate of absorption. Because of hyperemia, absorption will be faster from inflamed than from normal areas, unless the presence of edema lengthens the mean distance between capillaries and, thus, negates the effects of hyperemia on absorption.

Vasoconstriction may have a profound effect upon the rate of absorption. When a local effect of a drug is desired, as in local anesthesia, absorption away from the infiltered site may be impeded greatly by vasoconstrictors included in the preparation. Unwanted vasoconstriction sometimes may cause serious problems. For example, on World War II battlegrounds many wounded soldiers were given subcutaneous morphine without evident effect. As a result, injections were sometimes repeated more than once. When the patient was removed to the field hospital, toxic effects would occur suddenly. The explanation is that cold-induced vasoconstriction occurred in the field; when the patient was warmed in the hospital, vasodilation would result and the victim would be flooded with drug. Shock also contributes to the effect, since during shock the blood flow is diminished, and there also may be a superimposed vasoconstriction; repair of the shock condition then facilitates absorption.

Extravascularly injected molecules too large to pass through the capillary endothelium will, of necessity, enter the systemic circulation through the lymph. Thus, the lymph flow may be important to the absorption of a few drugs.

MOVEMENT—A number of factors combine so that movement at the site of injection increases the rate of absorption. In the intestine, segmental movements and peristalsis aid in dividing and dispersing the drug mass. The continual mixing of the chyme helps keep the concentration maximal at the mucosal surface. The pressures developed during segmentation and peristalsis also may favor a small amount of filtration. Movement at the site of hypodermic injection also favors absorption, since it tends to force the injected material through the tissue, increasing the surface area of drug mass and decreasing the mean distance to the capillaries. Movement also increases the flow of blood and lymph. The selection of a site for intramuscular injection may be determined by the amount of expected movement, according to whether the preparation is intended as a fast-acting or a repository preparation.

GASTRIC MOTILITY AND EMPTYING—The motility of the stomach is more important to the rate at which an orally administered drug is passed on to the small intestine than it is to the rate of absorption from the stomach itself, since for various reasons noted above, absorption from the stomach is usually of minor importance.

The average emptying time of the unloaded stomach is about 40 min, and the half-time is about 10 min, though it varies according to its contents, reflex, and psychological factors, and the action of certain autonomic drugs or disease. The effect of food to delay absorption is due, in part, to its action to prolong emptying time. The emptying time causes a delay in the absorption of drug, which may be unfavorable or favorable according to what is desired. In the case of therapy with antacids, gastric emptying is a nuisance, since it removes the antacid from the stomach where it is needed.

SOLUBILITY AND BINDING—The dissolution of drugs of low solubility is generally a slow process. Indeed, low solubility is the result of a low rate of departure of drug molecules from the undispersed phase. Furthermore, since the concentration around the drug mass is low, the concentration gradient from the site of deposition to the plasma is small, and the rate of diffusion is low, accordingly.

When it is desired that a drug have a prolonged action but not a high plasma concentration, a derivative of low solubility is often sought. The *insoluble* estolates and other esters of several steroids have durations of action of weeks because of the slow rates of absorption from the sites of injection. Insoluble salts or complexes of acidic or basic drugs also are employed as repository preparations; for example, the procaine salt of penicillin G has a low solubility and is used in a slow-release form of the antibiotic.

The solubility of certain macromolecules depends critically on the ionization of substituent groups. When they are amphiprotic, they are least soluble at their isoelectric pH. Insulin is normally soluble at the pH of the extracellular fluid, but by combining insulin with the right proportion of a basic protein, such as protamine, the isoelectric pH can be made to be approximately 7.4 from 5.1, and the complex can be used as a low-solubility, prolonged-action drug. For more details, see Chapter 77.

Some drugs may bind with natural substances at or near the site of application. The strongly ionized mucopolysaccharides in connective tissue, ground substance, and mucous secretions of the gut retard the absorption of a number of drugs, especially large cationic or polycationic molecules. In the gut, the binding is the least at low pH, which should favor absorption of large cations from the stomach; however, absorption from the stomach is slow (see above), so that the absorption of large cations occurs mainly in the upper duodenum where the pH is still relatively low. Pharmacologically inactive quaternary ammonium compounds sometimes are included in an oral preparation of a quaternary ammonium drug for the purpose of saturating the binding sites of mucin and other mucopolysaccharides and, thereby, enhancing the absorption of drug.

In addition to mucopolysaccharides in mucous secretions, food in the GI tract binds many drugs and slows absorption. Antacids, especially aluminum hydroxide plus other basic aluminum compounds and magnesium trisilicate, bind amine and ammonium drugs and interfere with absorption.

DONNAN EFFECT—The presence of a charged macromolecule on one side of a semipermeable membrane (impermeable to the macromolecule) will alter the concentration of permeant ionized particles according to the Donnan equilibrium. Accordingly, drug molecules of the same charge as the macromolecule will be constrained to the opposite side of the membrane. The presence of appropriately charged macromolecules

not only will influence the distribution of drug ions in accordance with the Donnan equation but also increase the rate of transfer of the drug across the membrane, because of mutual ionic repulsion. This effect is sometimes used to facilitate the absorption of ionizable drugs from the GI tract. The Donnan effect also operates to retard the absorption of drug ions of opposite charge; however, the mutual electrostatic attraction of a macromolecule and drug ion generally results in actual binding, which is more important than the Donnan effect.

VEHICLES AND ABSORPTION ADJUVANTS—Drugs that are to be applied topically to the skin and mucous membranes often are dissolved in vehicles that are thought to enhance penetration. For a long time it was thought that oleaginous vehicles promoted the absorption of lipid-soluble drugs. However, the role and effect of the vehicle has proven to be quite complex. In the skin at least five factors are involved:

1. The effect of the vehicle to alter the hydration of the keratin in the barrier layer.
2. The effect of the vehicle to promote or prevent the collection of sweat at the surface of the skin.
3. The partition coefficient of the drug in a vehicle-water system.
4. The permeability of the skin to the undissolved drug.
5. The permeability of the skin to the vehicle.

The effect of the vehicle to aid in the access of the drug to the hair follicles and sebaceous glands also may be involved, although in man the follicles and glands are probably ordinarily of minor importance to absorption.

A layer of oleaginous material over the skin prevents the evaporation of water, so that the stratum corneum may become macerated and more permeable to drugs. In dermatology it is sometimes the practice to wrap the site of application with plastic wrap or some other waterproof material for the purpose of increasing the maceration of the stratum corneum. However, the layer of perspiration that forms under an occlusive vehicle may become a barrier to the movement of lipid-soluble drugs from the vehicle to the skin, but it may facilitate the movement of water-soluble drugs. Conversely, polyethylene glycol vehicles remove the perspiration and dehydrate the barrier, which decreases the permeability to drugs; such vehicles remove the aqueous medium through which water-soluble drugs may pass down into the stratum corneum but at the same time facilitate the transfer of lipid-soluble drugs from the vehicle to the skin.

Even in the absence of a vehicle, it is not clear what physicochemical properties of a drug favor cutaneous penetration, high lipid-solubility being a prerequisite, according to some authorities, and an ether-water partition coefficient of approximately one, according to others. Yet, the penetration of ethanol and dibromomethane are nearly equal, and other such enigmas exist. It is not surprising, then, that the effects of vehicles are not altogether predictable.

A general statement might be made that if a drug is quite soluble in a poorly absorbed vehicle, the vehicle will retard the movement of the drug into the skin. For example, salicylic acid is 100 times as permeant when absorbed from water than from polyethylene glycol, and pentanol is five times as permeant from water as from olive oil. Yet, ethanol penetrates five times faster from olive oil than from either water or ethanol, all of which denies the trustworthiness of generalizations about vehicles.

For several decades there has been much interest in certain highly dielectric, aprotic solvents, especially dimethyl sulfoxide (DMSO). Such substances generally prove to be excellent solvents for both water- and lipid-soluble compounds and for some compounds not soluble in either water or lipid solvents. The extraordinary solvent properties probably are due to a high polarizability and van der Waals bonding capacity, a high degree of polarization (dipole moment), and a lack of association through hydrogen bonding. As a vehicle, DMSO greatly facilitates the permeation of the skin and other biological membranes by numerous drugs, including such large molecules as insulin. The mechanism is understood poorly. Such vehicles have a potential for many important uses, but they are at present only experimental, pending continuing investigations on toxicity.

From time to time, a claim is made that a new ingredient of a tablet or elixir enhances the absorption of a drug, and a comparison of plasma levels of the old and new preparations seems to support the claim. Upon further investigation, however, it may be revealed that the new so-called absorption adjuvant is replacing an ingredient that previously bound the drug or delayed its absorption; thus, the new *adjuvant* is not an adjuvant but rather it is only a nondeterrent.

OTHER FACTORS—A number of other, less well-defined factors affect the absorption of drugs, some of which may operate, in part, through factors already cited above. Disease or injury has a considerable effect upon absorption. For example, debridement of the stratum corneum increases the permeability to topical agents, meningitis increases the permeability of the blood-brain barrier, biliary insufficiency decreases the absorption of lipid-soluble substances from the intestine, and acid-base disturbances can affect the absorption of weak acids or bases. Certain drugs, such as ouabain, that affect active transport processes may interfere with the absorption of certain other drugs. The condition of the *ground substance,* or *intracellular cement,* probably bears on the absorption of certain types of molecules. Hyaluronidase, which depolymerizes the mucopolysaccharide ground substance, can be demonstrated to facilitate the absorption of some, but not all, drugs from subcutaneous sites.

Drug Disposition

The term *drug disposition* is used here to include all processes that tend to lower the plasma concentration of drug, as opposed to drug absorption, which elevates the plasma level. Consequently, the distribution of drugs to the various tissues is considered under *Disposition*. Some authors use the term disposition synonymously with elimination, that is, to include only those processes that decrease the amount of drug in the body. In the present context, disposition comprises three categories of processes: distribution, biotransformation, and excretion.

DISTRIBUTION, BIOTRANSFORMATION, AND EXCRETION

The term *distribution* denotes the partitioning of a drug among the numerous locations where a drug may be contained within the body. *Biotransformations* are the alterations in the chemical structure of a drug that are imposed upon it by the life processes. *Excretion* is, in a sense, the converse of absorption, namely, the transportation of the drug or its products out of the body. The term applies whether or not special organs of excretion are involved.

Distribution

The body may be considered to comprise a number of *compartments:* enteric (GI), plasma, interstitial, cerebrospinal fluid, bile, glandular secretions, urine, storage vesicles, cytoplasm or intracellular space, etc. Some of these *compartments,* such as urine and secretions, are open-ended, but since their contents relate to those in the closed compartments, they also must be included.

At first thought, it may seem that if a drug was distributed passively (ie, by simple diffusion) and the plasma concentration could be maintained at a steady level, the concentration of a drug in the water in all compartments ought to become equal. It is true that some substances, such as ethanol and antipyrine, are distributed nearly equally throughout the body water, but they are more the exception than the rule. Such substances are mainly small, uncharged, nondissociable, highly water-soluble molecules.

The condition of small size and high water solubility allows passage through the pores without the necessity of carrier or active transport. Small size also places a limit on van der Waals binding energy and configurational complementariness, so that binding to proteins in plasma, or cells, is slight. The presence of a charge on a drug molecule makes for unequal distribution across charged membranes, in accordance with the Donnan distribution (see below). Dissociability causes unequal distribution when there is a pH differential between compartments, as discussed under *The pH Partition Principle* (see below). Thus, even if a drug is distributed passively, its distribution may be uneven throughout the body. When active transport into, or rapid biotransformation occurs within, some compartments, uneven distribution also is inevitable.

THE pH PARTITION PRINCIPLE—An important consequence of nonionic diffusion is that a difference in pH between two compartments will have an important influence upon the partitioning of a weakly acidic or basic drug between those compartments. The partition is such that the un-ionized form of the drug has the same concentration in both compartments, since it is the form that is freely diffusible; the ionized form in each compartment will have the concentration that is determined by the pH in that compartment, the pK and the concentration of the un-ionized form. The governing effect of pH and pK on the partition is known as the *pH partition principle*.

To illustrate the principle, consider the partition of salicylic acid between the gastric juice and the interior of a gastric mucosal cell. Assume the pH of the gastric juice to be 1, which it occasionally becomes. The pK_a of salicylic acid is 3 (Martin[10] provides one source of pK values of drugs). With the Henderson-Hasselbalch equation (see Chapter 17) it may be calculated that the drug is only 1% ionized at pH 1. (The relationship of ionization and partition to pH and pK has been formulated in several different ways, but the student may calculate the concentrations from simple mass law equations. More sophisticated calculations and reviews of this subject are available.[6,11–16]) The intracellular pH of most cells is about 7. Assuming the pH of the mucosal cell to be the same, it may be calculated that salicylic acid will be 99.99% ionized within the cells. Since the concentration of the un-ionized form is theoretically the same in both gastric juice and mucosal cells, it follows that the total concentration of the drug (ionized + un-ionized) within the mucosal cell will be 10,000 times greater than that in gastric juice. This is illustrated in Figure 57-13. Such a relatively high intracellular concentration can have important osmotic and toxicological consequences.

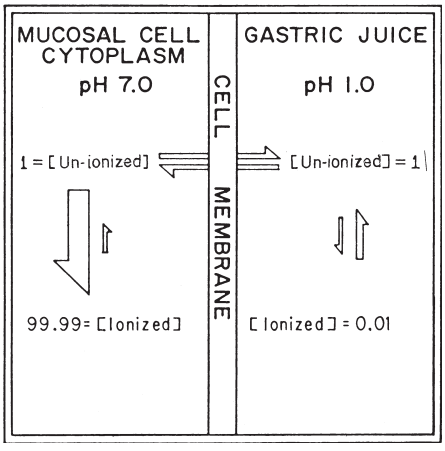


Figure 57-13. Hypothetical partition of salicylic acid between gastric juice and the cytoplasm of a gastric mucosal cell. It is assumed that the ionized form cannot pass through the cell membrane. The intragastric concentration of salicylic acid is arranged arbitrarily to provide unit concentration of the un-ionized form. *Bracketed values,* concentration; *arrows,* relative size depicts the direction in which dissociation-association is favored at equilibrium.

Had the drug been a weak base instead of an acid, the high concentration would have been in the gastric juice. In the small intestine, where the pH may range from 7.5 to 8.1, the partition of a weak acid or base will be the reverse of that in the stomach, but the concentration differential will be lower, because the pH differential from lumen to mucosal cells, etc, will be lower. The reversal of partition as the drug moves from the stomach to the small intestine accounts for the phenomenon that some drugs may be absorbed from one GI segment and returned to another. The weak base atropine is absorbed from the small intestine, but because of pH partition, it is *secreted* into the gastric juice.

The pH partition of drugs has never been demonstrated to be as marked as that illustrated in Figure 57-13 and in the text. Not only do many drug ions probably pass through the pores of the membrane to a significant extent, but also some may pass through the lipid phase, as explained above for the morphinans and mecamylamine. Furthermore, ion-pair formation in carrier transport also bypasses nonionic diffusion. All processes that tend toward an equal distribution of drugs across membranes and among compartments will cause further deviations from theoretical predictions of pH partition.

ELECTROCHEMICAL AND DONNAN DISTRIBUTION—A drug ion may be distributed passively across a membrane in accordance with the membrane potential, the charge on the drug ion, and the Donnan effect. The relationship of the membrane potential to the passive distribution of ions is expressed quantitatively by the Nernst equation (Eq 7) and already has been discussed. Barring active transport, pH partition, and binding, the drug will be said to be distributed according to the electrical gradient or to its *equilibrium* potential. If the membrane potential is 90 mV, the concentration of a univalent cation will be 30 times as high within the cell as without; if the drug cation is divalent, the ratio will be 890. The distribution of anions would be just the reverse. If the membrane potential is but 9 mV, the ratio for a univalent cation will be only 1.4 and for a divalent cation only 2.0. It thus can be seen how important membrane potential may be to the distribution of ionized drugs.

It was pointed out under *Membrane Potentials,* that large potentials derive from active transport of ions but that small potentials may result from Donnan distribution. Donnan membrane theory is discussed in Chapter 20. According to the theory, the ratio of intracellular/extracellular concentration of a permeant univalent anion is equal to the ratio of extracellular/intracellular concentration of a permeant univalent cation. A more general mathematical expression that includes ions of any valence is

$$\left(\frac{A_i}{A_e}\right)^{1/Z_a} = \left(\frac{C_e}{C_i}\right)^{1/Z_c} = r \qquad (8)$$

where A_i is the intracellular and A_e the extracellular concentration of anion, Z_c is the valence of cation, Z_a is the valence of anion, C_i is the intracellular and C_e the extracellular concentration of cation, and r is the Donnan factor. The value of r depends upon the average molecular weight and valence of the macromolecules (mostly protein) within the cell and the intracellular and extracellular volumes. Since the macromolecules within the cell are charged negatively, the cation concentration will be higher within the cell; that is, $C_i > C_e$. Since a Donnan distribution results in a membrane potential, the distribution of drug ion also will be in keeping with the membrane potential.

The Donnan distribution also applies to the distribution of a charged drug between the plasma and interstitial compartment, because of the presence of anionic proteins in the plasma. Equation 8 applies by changing the subscript i to p, for plasma, and e to i, for interstitial. The Donnan factor, r, for plasma–interstitial space partition is about 1.05:1.

BINDING AND STORAGE—Drugs frequently are bound to plasma proteins (especially albumin), interstitial substances, intracellular constituents, and bone and cartilage. If binding is

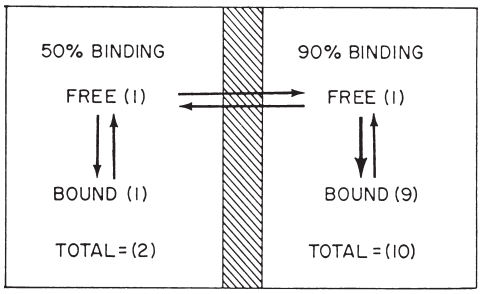


Figure 57-14. Distribution of a drug between two compartments in which the degrees of binding to protein differ. The percentage of binding is indicated. Only the unbound drug can pass through the membrane. *Bracketed values:* concentration. (From Schanker LS. *Pharmacol Rev* 1961; 14:501.)

extensive and firm, it will have a considerable impact upon the distribution, excretion, and sojourn of the drug in the body. Obviously, a drug that is bound to a protein or any other macromolecule will not pass through the membrane in the bound form; only the unbound form can negotiate among the various compartments.

The partition among compartments is determined by the binding capacity and binding constant in each compartment. As long as the binding capacity exceeds the quantity of drug in the compartment, the following equation generally applies:

$$\log D_b = \log K + a \log D_f$$

where D_b is the concentration of bound drug, D_f is the concentration of free drug, and a and K are constants characteristic of the drug and binding macromolecule. The equation is that of a Freundlich isotherm. As the binding capacity is approached, the relationship no longer holds. For a nondissociable drug at equilibrium, D_f will be the same in all communicating compartments, so that it would be possible to calculate the partition if K and a are known for each compartment. Except for plasma, the values of K and a are generally unknown, but the percentage bound is often known.

From the percentage bound, the partition also can be calculated, as in Figure 57-14.[12] However, the logarithmic relationships shown in Equation 9 serve as a reminder that the percentage bound changes with the concentration, so that the partition will vary with the dose. If the drug is a weak acid or base, the un-ionized free form negotiates among the compartments, but the ionized form is often the more firmly bound, and calculations must take into account the dissociation constant and the different Ks and as of the ionized and un-ionized forms.

It is misbelieved commonly that binding in the plasma interferes with the activity of a drug and the intracellular binding in a responsive cell increases activity or toxicity. Both binding in plasma and in the tissues decreases the concentration of free drug, but this is easily remedied by adjusting the dose to give a sufficient concentration for pharmacological activity. The distribution and activity of the free form are not affected by binding. The principal effect of binding is to increase the initial dose requirement for the drug and create a reservoir of drug from which the drug may be withdrawn as the free form is excreted or metabolized. However, if the binding is extremely firm and release is slow, the rate of release may not be enough to sustain the free form at a level sufficient for pharmacological activity; in such instances the bound drug cannot be considered a reserve.

The effect of binding upon the sojourn of a drug may be considerable. For example, quinacrine, which may be concentrated in the liver to as much as several thousand times the concentration in plasma, may remain in the body for months. Some iodine-containing, radiopaque, diagnostic agents are bound strongly to plasma protein and may remain in the plasma for as long as 2 yr. In pathological conditions, such as nephrosis, diabetes, or cirrhosis, in which plasma protein levels may be decreased, the plasma protein binding, loading dose, and duration of action all may be decreased.

If a drug is bound to a functional macromolecule, binding may relate to pharmacological activity and toxicity, providing that the binding is at a critical center of the macromolecule. The binding by nucleic acids of certain antimalarials, such as quinacrine, undoubtedly contributes to the parasiticidal actions as well as to toxicity.

Most drugs are bound to proteins by relatively weak forces, such as van der Waals (London, Keesom, or Debye) forces, or hydrogen or ionic bonds. Consequently, binding constants generally are small, and binding is usually readily reversible. The larger the molecule, the greater the van der Waals bonding, so that large drug molecules are more likely to be bound strongly than are small ones.

Just as shape and the nature of functional groups are important to drug-receptor combination, so they also are to binding. Drugs of similar shape and/or chemical affinities may bind at the same sites on a binding protein and hence compete with one another. For example, phenylbutazone displaces warfarin from human plasma albumin, which may cause an increase in the anticoagulant effect of warfarin. Some drugs also may displace protein-bound endogenous constituents. For example, sulfisoxazole displaces bilirubin from plasma proteins; in infants with kernicterus the freed bilirubin floods the CNS and causes sometimes fatal toxicity.

Depending on the lipid-water partition coefficient, a drug may be taken up into fatty tissue. The ratio of concentration in fat to that in the plasma, will not be the same as dictated by the partition coefficient, because of the content of water and nonlipids in adipose tissue, and because electrolytes and other solutes alter the dielectric constant and hence solubilities from those of pure water. Lipoproteins and even nonpolar substituents on plasma proteins also take up lipid-soluble molecules, so that solubility in plasma can be considerably higher than that in water. The relatively high solubility of ether in plasma makes plasma a pool for ether, the filling of which delays the onset of anesthesia. However, ether and other volatile anesthetics are taken up gradually into the adipose tissue, which acts as a store of the anesthetic. The longer the anesthetic is administered, the greater the store, and the longer it takes for anesthesia to terminate when inhalation has been discontinued.

Another notable substance that is taken up readily into fat is thiopental. Even though there is a high solubility of this barbiturate in fat, the low rate of blood flow in fat limits the rate of uptake. Because the blood flow in the brain is very high, thiopental rapidly enters brain tissue. However, it soon equilibrates with the other tissues, and the brain concentration falls as that in the other tissues (eg, muscle or liver) increases. As the brain concentration falls, anesthesia ceases. Gradually, the fat accumulates the drug at the expense of other compartments. The gradual entry of thiopental into fat at the expense of plasma, muscle, or liver is illustrated in Figure 57-15.

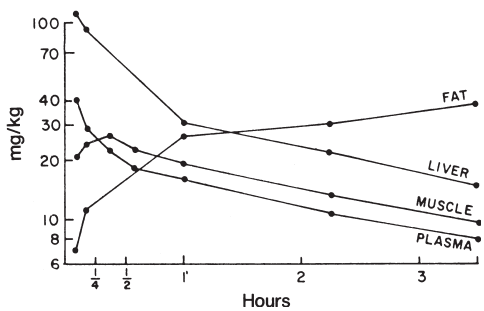


Figure 57-15. Predisposition of thiopental for fat; 25 mg/kg was given to a dog. After a brief sojourn in the more vascular tissues, thiopental gradually transfers to fat, where the lipid-soluble drug dissolves in fat droplets. (From Brodie BB, Hogben CA. *J Pharm Pharmacol* 1957; 9:345.)

NONEQUILIBRIUM AND REDISTRIBUTION—Thus far, the distribution of drugs has been discussed mainly as though equilibrium or steady-state conditions exist after a drug is absorbed and distributed. However, since most drugs are administered at intervals and the body content of drug rises and falls with absorption and biotransformation-excretion, neither a true equilibrium among the body compartments nor a steady state exist.

The term equilibrium is used misleadingly to describe the conditions that exist when the plasma concentration and the concentration in a tissue are equal, as exemplified at the point of intersection of the curves for plasma and muscle or plasma and fat in Figure 57-15. But such *equilibrium* with fat occurs much later than *equilibrium* with muscle, so that no true equilibrium really exists among all the compartments. Furthermore, the crossover point for plasma and any one tissue is not necessarily an equilibrium point, because the rates of ingress and egress from the tissue are not necessarily equal when the internal and external concentrations are equal, since there are numerous factors that make for unequal distribution (pH partition, Donnan effect, electrochemical distribution, active transport, binding, etc).

A study of Figure 57-15 shows that the distribution of thiopental continually changed during the 3.5 hr of observation. At the end of the period, the content in fat was still increasing, while that in each of the other compartments was decreasing. This time-dependent shift in partition is called *redistribution*. Eventually, the content in fat would have reached a peak, which would represent as nearly a true equilibrium point as could be achieved in the dynamic situation where biotransformation and a slight amount of excretion of the drug was taking place. Once the concentration in the fat had reached its peak, its content would have declined in parallel with that in the other tissues, and the partition among the compartments would have remained essentially constant. Redistribution, then, takes place only until the concentration in the slowest-filling compartment reaches its peak, so long as the kinetics of elimination are constant.

An index of distribution known as the *volume of distribution* (amount of drug in the body divided by plasma concentration) is of considerable usefulness in pharmacokinetics but is of limited value in defining the way in which a drug is partitioned in the body.

The word *space* often is used synonymously with volume of distribution. It is employed especially when the distributed substance has a volume of distribution that is essentially identical to a physical real space or body compartment. *N*-acetyl-4-aminoantipyrine is distributed evenly throughout the total body water and is not bound to proteins or other tissue constituents. Thus, the acetylaminoantipyrine space, or volume of distribution, coincides with that of total body water. Inulin, sucrose, sulfate, and a number of other substances essentially are confined to extracellular water, so that an inulin space, for example, measures the extracellular fluid volume. Evans blue is confined to the plasma, so that the Evans blue space is the plasma volume. Such space measurements with standard space indicators are a necessary part of studies on the distribution of drugs, since it is desirable to compare the volume of distribution of a drug with the physiological spaces.

Biotransformations

Most drugs are acted upon by enzymes in the body and converted to metabolic derivatives called metabolites. The process of conversion is called *biotransformation*. Metabolites are usually more polar and less lipid-soluble than the parent drug because of the introduction of oxygen into the molecule, hydrolysis to yield more highly polar groups, or conjugation with a highly polar substance. As a consequence, metabolites often show less penetration into tissues and less renal tubular resorption than the parent drug, in accordance with the principle of the low penetration of polar and high penetration of lipid-soluble substances. For similar reasons, metabolites, particularly conjugates, are usually less active than the parent drug and often inactive. Even if they are appreciably active, they generally are excreted more rapidly. Therefore, the usual net effect of biotransformation may be said to be one of *inactivation* or *detoxication*.

There are, however, numerous examples in which biotransformation does not result in inactivation.

There are also examples in which the parent drug has little or no activity of its own but is converted to an active metabolite: parathion, malathion, and certain other anticholinesterases require metabolic activation; inactive chloroguanide is converted to an active triazine derivative; phenylbutazone is hydroxylated to the antirheumatic hydroxyphenylbutazone; inactive pentavalent arsenicals are reduced to their active trivalent metabolites, and there are other examples of an activating biotransformation.

When a delayed or prolonged response to a drug is desired or an unpleasant taste or local reaction is to be avoided, it is a common pharmaceutical practice to prepare an inactive or nonoffending precursor, such that the active form may be generated in the body. This practice has been termed *drug latentiation*. Chloramphenicol palmitate, dichloralphenazone, and the estolates of various steroid hormones are examples of deliberately latentiated drugs. Because inactive metabolites do not always result from biotransformation, the term detoxication should not be used as a synonym for biotransformation.

Biotransformations take place principally in the liver, although the kidney, skeletal muscle, intestine, or even plasma may be important sites of the enzymatic attack of some drugs. Biotransformations in plasma are mostly hydrolytic.

ENDOPLASMIC RETICULUM AND MICROSOMAL SYSTEM—Many biotransformations in the liver occur in the *endoplasmic reticulum*. The endoplasmic reticulum is a tubular system that courses through the interior of the cell but also appears to communicate with the interstitial space, and its membrane is continuous with the cell membrane. Some of the reticulum is lined with ribonucleoprotein particles, called ribosomes, which are engaged in protein synthesis; this is the *rough* endoplasmic reticulum. The smooth endoplasmic reticulum lacks such a granular appearance. The endoplasmic reticulum is invested heavily with numerous enzymes, which biotransform many drugs and some endogenous substances.

When a broken-cell homogenate of the liver is prepared, the reticulum becomes fragmented, and the fragments form vesicular structures called *microsomes*. Although the microsomes are artifacts, it is often the practice to refer to drug metabolism as occurring in microsomes rather than in the endoplasmic reticulum.

The microsomal system is peculiar in that both oxidations and reductions usually require the reducing cofactor, reduced nicotinamide adenine dinucleotide phosphate (NADPH). This is because microsomal oxidations proceed by way of the introduction of oxygen rather than by dehydrogenation, and NADPH is essential to reduce one of the atoms of oxygen. The drug first binds to an oxidized cytochrome P450. The drug-cytochrome complex then is reduced by NADPH–cytochrome P450 reductase; the reduced complex then combines with oxygen, after which the metabolite is released and oxidized cytochrome P450 is regenerated. Cytochrome P450 is a generic term for a superfamily of enzymes.[17]

The general designation of the cytochromes P450 is *CYP* followed by number (the family) and letter (the subfamily) subdivisions. The classification is based on amino acid sequence homology. To belong to the same family, the homology must be greater than 40% and to the same subfamily greater than 59%. The form is indicated by a number that is based upon the chronological discovery order. The major human forms involved in drug metabolism are CYP1A1 and CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9/10, CYP2C18/19, CYP2D6, CYP2E1, CYP3A4, CYP3A5, and CYP3A7. In concentration, CYP3As comprise 40% of the liver P450, CYP2Cs comprise 25%, and CYP1A2 about 15%. Despite its limited concentration (2%), CYP2D6 metabolizes about one-fourth of currently used drugs and is widely tested for because of a genetic polymor-

phism in which 5 to 10% of the population are poor metabolizers. The different isozymes present in humans, together with which drugs they metabolize, are of increasing importance in understanding drug interactions and toxicities and individual responses to standardized doses.

In addition to cytochrome P450, the endoplasmic reticulum contains flavoprotein monooxygenases, which are also responsible for the oxidative metabolism of drugs. The mechanism of oxidation differs from that of cytochrome P450, and their substrate (any drug containing a nucleophilic heteroatom) selectivity is much less. *FMO3* is the major human liver form.

Some of the enzymes of the microsomal system are quite easily *induced;* that is, a drug may increase considerably the activity of the enzyme by increasing the biosynthesis of the enzyme. An increase in the amount of endoplasmic reticulum sometimes occurs concomitantly with enzyme induction.

The mechanism of induction is best documented for polycyclic aromatic hydrocarbon (Ah)-type inducers but is thought to be similar for all agents; however, it involves different receptors, which interact with different regulatory elements on the DNA (Fig 57-16). The cytosol contains proteins that have a high affinity for the inducing agents. In normal drug therapy, the drug (D) enters the liver cell and, if adequately metabolized, is discharged as metabolites. Inefficient clearance from the cell, possibly due to high dosage, results in accumulation (ie, excess), and some is able to bind to the protein, which has a high affinity for the accumulating drug. When the inducing agent binds to its receptor, there is a conformational change (for an Ah receptor, chaperone proteins are displaced) allowing the receptor-inducer complex to translocate into the nucleus, link with additional nuclear factors, and initiate the transcription of mRNA to a limited number of proteins, by binding to DNA regions termed a drug-response element (DRE) (xenobiotic response element for the Ah receptor complex) that activate gene transcription. (For polycyclic aromatic hydrocarbons, the activated genes including specific isozymes of cytochrome P450, glutathione S-transferase, and UDP-glucuronosyltransferase.) These mRNA molecules move out of the nucleus and are translated into new proteins on the ribosomes attached to the endoplasmic reticulum.

The drug-metabolizing enzymes differ in their ability to be induced. For cytochrome P450s, CYP1A2 is induced preferentially by polycyclic aromatic hydrocarbons and other chemicals contained in cigarette smoke and charcoal-broiled meats, as well as by components in cruciferous vegetables. CYP2A6 is induced by barbiturates as are CYP2C9 and CYP3A4. CYP2C9, CYP2C19, and CYP3A4 are all induced by rifampicin, but CYP3A4 is additionally induced by many drugs including carbamazepine, phenytoin, glucocorticoids (dexamethasone), clotrimazole, sulfinpyrazone, and macrolide antibiotics such as troleandomycin. CYP2E1 can be induced by ethanol and isoniazid. There are no known inducers of CYP2D6.

Treatment of an experimental subject with phenobarbital will increase the rate of metabolism of phenobarbital, which necessitates larger and more frequent doses of the drug to maintain a constant sedative effect. Moreover, phenobarbital may induce an increased metabolism of some other, but not all, barbiturates as well as some unrelated drugs, such as strychnine and warfarin. Oddly, warfarin does not induce its own biotransformation readily.

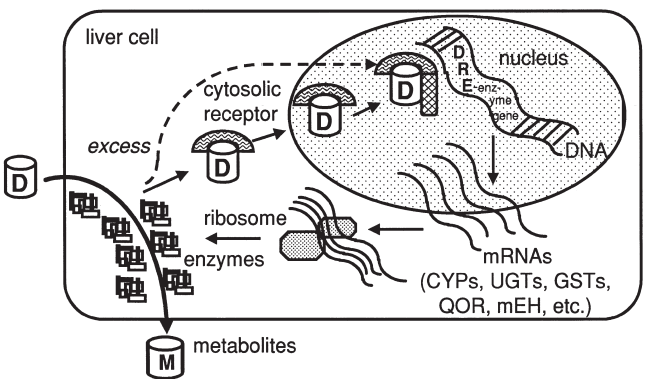


Figure 57-16.

Induction may create therapeutic problems. For example, the use of phenobarbital during treatment with warfarin increases the dose requirement for warfarin. If the physician is unaware of this interaction and fails to increase the dose, the patient may suffer a thrombotic episode. If the dose of warfarin has been increased and the phenobarbital is then discontinued, the rate of metabolism of warfarin may drop to its previous level, so that the patient is overdosed, with hemorrhagic consequences. Some drugs inhibit rather than induce the microsomal enzymes, which reduces the dose requirement and may lead to toxicity. Cimetidine is an example of a drug that inhibits the hepatic metabolism of a number of other drugs.

The activity of the microsomal biotransformation enzymes is affected by many factors other than the presence of drugs. Age, sex, nutritional states, pathological conditions, and genetic factors are among the influences that have been identified. Age, particularly, has received considerable attention. Infants have a poorly developed microsomal biotransformation system, which accounts for the low dose requirement for morphine and also explains the high toxicity of chloramphenicol in infants.

The activity and selectivity of the microsomal biotransformation system varies greatly from species to species, so that care must be exercised in extrapolating experimental findings in laboratory animals to man.

TYPES OF BIOTRANSFORMATIONS—Biotransformations may be *degradative,* wherein the drug molecule is diminished to a smaller structure, or *synthetic,* wherein one or more atoms or groups may be added to the molecule. Very few drugs are degraded completely. However, it is more useful to categorize biotransformations with respect to *metabolic* (nonconjugative) biotransformations and conjugative biotransformations. The former is called Phase I and the latter, Phase II. In Phase I, pharmacodynamic activity may be lost; however, active and chemically reactive intermediates also may be generated. The polarity of the molecule may or may not be increased sufficiently to increase excretion markedly. In Phase II, metabolites from Phase I may be conjugated, and sometimes the original drug may be conjugated, thus bypassing Phase I. Phase II generates metabolites of high polarity, which are excreted readily.

Biotransformations may be placed into four main categories: (1) oxidation, (2) reduction, (3) hydrolysis, and (4) conjugation. Oxidation, reduction, and hydrolysis comprise Phase I. Conjugation comprises Phase II.

Oxidation—Oxidation is more common than any other type of biotransformation. Oxidations that occur primarily in the liver microsomal system include side-chain hydroxylation; aromatic hydroxylation; deamination (which is oxidative and results in the intermediate formation of RCHO); *N-, O-,* and *S*-dealkylation (which probably involves hydroxylation of the alkyl group followed by oxidation to the aldehyde); and sulfoxide formation.

Oxidations that occur elsewhere, other than the microsomes, are generally dehydrogenations followed by the addition of oxygen or water. Examples are the oxidation of alcohols by alcohol dehydrogenase, the oxidation of aldehyde by aldehyde dehydrogenase, and the deamination of monoamines by monoamine oxidase and diamines by diamine oxidase.

Reduction—Reductions are relatively uncommon. They mainly occur in liver microsomes, but they occasionally take place in other tissues. Examples are the reduction of nitro and nitroso groups (as in chloramphenicol, nitroglycerin, and organic nitrites), of the azo group (as in prontosil), and of certain aldehydes to the corresponding alcohols.

Hydrolysis—Hydrolysis is a common biotransformation among esters and amides. Esterases are located in many structures besides the microsomes. For example, cholinesterases are found in plasma, erythrocytes, liver, nerve terminals, junctional interstices, and postjunctional structures, and procaine esterases are found in plasma. Various phosphatases and sulfatases also are distributed widely in tissues and plasma, although few drugs are appropriate substrates. The hydrolytic deamidation of meperidine occurs primarily in the hepatic microsomes.

The hydrolysis of epoxides, often generated by cytochrome P450 oxidations, to form dihydrodiols is an important detoxification reaction.

Desulfuration, in which oxygen may replace sulfur, takes place in the liver. Thiopental is converted in part to pentobarbital by desulfuration, and parathion is transformed to paraoxon.

Dehalogenation of certain insecticides and various halogenated hydrocarbons may take place, principally in the liver but not in the microsomes.

Conjugation—A large number of drugs, or their metabolites, are conjugated. Conjugation is the biosynthetic process of combining a chemical compound with a highly polar and water-soluble natural substance to yield a water-soluble, usually inactive, product. Conjugations generally involve either esterification, amidation, mixed anhydride formation, hemiacetal formation, or etherization.

Glucuronic acid is the most frequent partner to the drug in conjugation. Actually, the drug reacts with uridine diphosphoglucuronic acid rather than with simple glucuronic acid. The drug or drug metabolite combines at the number 1 carbon (aldehyde end) and not at the carboxyl end of glucuronic acid. The hydroxyl group of an alcohol or a phenol attacks the number 1 carbon of the pyran ring to replace uridine diphosphate. The product is a hemiacetal-like derivative. Since the product is not an ester, the term *glucuronide* is appropriate. Rarely, thiols and amines may form analogous glucuronides.

Carboxyl compounds form esters, appropriately called glucuronates, in replacing the uridine diphosphate. *Sulfuric acid* is also a frequent conjugant, especially with phenols and to a lesser extent with simple alcohols. The sulfurated product is called an *ethereal sulfate*.

Occasionally sulfuric acid conjugates with aromatic amines to form *sulfamates*. *Phosphoric acid* also conjugates with phenols and aromatic amines. The conjugation of benzoic acid with glycine to yield hippuric acid is a classical example of an *amidation* conjugative process.

Many electrophilic compounds conjugate with the nucleophilic tripeptide, glutathione. Through a series of enzymatic reactions, the γ-glutamyl and glycyl residues are removed, the remaining cysteine conjugate is *N*-acetylated, and the product spontaneously dehydrates to form a mercapturic acid.

Amidations with amino acids are less frequent than *acetylation,* partly because few drugs are carboxylic compounds. Aromatic amines and occasionally aliphatic amines or heterocyclic nitrogen frequently are acetylated. Acetyl-CoA is the biological reagent rather than acetic acid itself. Unlike most other conjugates, the acetylate (amide) is usually less water-soluble than the parent compound. The acetylation of the para-amino group of the sulfonamides is a prime example of this type of conjugation.

Although most conjugations occur in the liver, some occur in the kidney or in other tissues.

Many amines, especially derivatives of β-phenylethylamine and heterocyclic compounds, are methylated in the body. The products are usually biologically active, sometimes more so than the parent compound. *N-Methylation* may occur in the cytoplasm of the liver and elsewhere, especially in chromaffin tissue in the case of phenylethylamines.

Phenolic compounds may be *O*-methylated. *O-Methylation* is the principal route of biotransformation of catecholamines such as epinephrine and norepinephrine, the methyl group being introduced on the *meta*-hydroxy substituent. Both *N*- and *O*-methylation require *S*-adenosylmethionine.

All the drug conjugation reactions are catalyzed by specialized enzymes present in multiple forms. Glucuronidation is catalyzed by UDP-glucuronosyltransferases, *UGTs,* located in the endoplasmic reticulum. UGTs are classified in two major classes, UGT1As and UGT2Bs, based on amino acid homology, but the two classes also differ in substrate selectivity, with UGT1As preferring planar drugs and UGT2Bs preferring bulkier molecules. As with cytochrome P450s, these enzymes are inducible, and the two classes differ in their response to various drugs and other chemicals.

Sulfation is catalyzed by sulfotransferases, *SULTs,* located in the cytoplasm. The many isozymes exhibit substrate selectivity, and some differ in thermal stability. Unlike most major drug-metabolizing enzymes, SULTs are refractory to induction by drugs.

Glutathione conjugations are catalyzed by glutathione-*S*-transferases, *GSTs,* also located in the cytoplasm. The multiple isozymes are designated into four major classes: alpha, mu, pi and theta. The isozymes have relatively low substrate (electrophile) selectivity. Methylation reactions are catalyzed by cytoplasmic *O*-, *N*-, and *S*-methyltransferases, and each exists in multiple forms.

Acetylation is catalyzed by cytoplasmic *N*-acetyltransferases, *NAT*1, and in the liver, *NAT*2. NAT2 exhibits a genetic polymorphism, giving *fast and slow* acetylator phenotypes with differing incidences in various populations (slow is high in Middle Eastern, low in Asian).

Excretion

Some drugs are not biotransformed in the body. Others may be biotransformed, but their products still remain to be eliminated. It follows that excretion is involved in the elimination of all drugs and/or their metabolites. Although the kidney is the most important organ of excretion, some substances are excreted in bile, sweat, saliva, or gastric juice or from the lungs.

RENAL EXCRETION—The excretory unit of the kidney is called the *nephron* (Fig 57-17). There are several million nephrons in the human kidney. The nephron is essentially a filter funnel, called *Bowman's capsule*, with a long stem, called a *renal tubule*. It also is recognized now that the collecting duct is functionally a part of the nephron. The *blood vessels* that invest the capsule and the tubule are also an essential part of the nephron.

Bowman's capsule is packed with a tuft of branching interconnected capillaries *(glomerular tuft),* which provide a large surface area of capillary endothelium *(filter paper)* through which fluid and small molecules may filter into the capsule and begin passage down the tubule. The glomerular tuft, together with Bowman's capsule, constitute the *glomerulus*. The glomerular capillary endothelium and the supporting layer of Bowman's capsule have channels ranging upward to 40 Å. Consequently, all unbound crystalloid solutes in plasma, and even a little albumin, pass or are forced by pressure into the glomerular filtrate.

The postglomerular vessels, which lie close to the tubules, are critically important to renal function in that substances resorbed from the filtrate by the tubule are returned to the blood along these vessels. The tubule is not straight but rather first

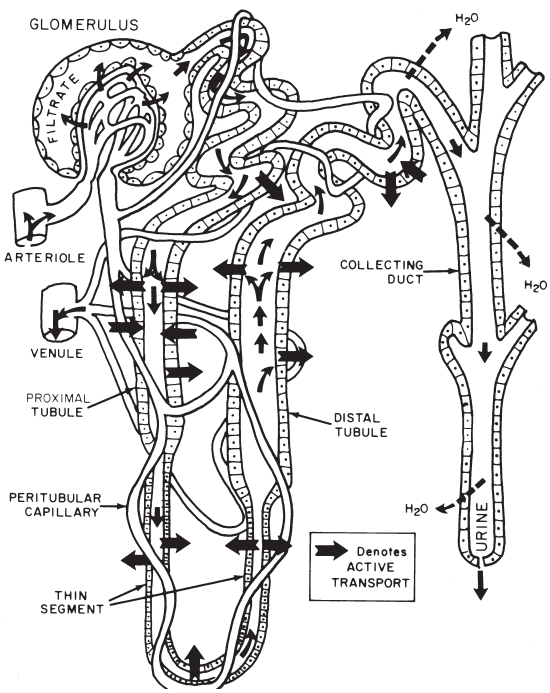


Figure 57-17. Diagram of a mammalian nephron. Note how the lower loops of the postglomerular capillaries course downward and double back along with the tubule. This allows countercurrent distribution to maintain hyperosmolar urine within the thin segment.

makes a number of convolutions (called a *proximal convoluted tubule*), then courses down and back up a long loop (called the *loop of Henle*), makes more convolutions (the *distal convoluted tubule*) and finally joins the collecting duct. The loop of Henle is divided into a *proximal (descending) tubule*, a thin segment and a *distal (ascending) tubule*.

As the glomerular filtrate passes through the proximal tubule, some solute may be resorbed *(tubular resorption)* through the tubular epithelium and returned to the blood. Resorption occurs in part by passive diffusion and in part by active transport, especially with sodium and glucose. Chloride follows sodium obligatorily.

In the proximal region, the tubule is quite permeable to water, so that resorbed solutes are accompanied by enough water to keep the resorbate isotonic. Consequently, although the filtrate becomes diminished in volume by approximately 80% in the proximal tubule, it is not concentrated.

Some *acidification* occurs in the proximal tubule as the result of carbonic anhydrase activity in the tubule cells and the diffusion of hydronium ions into the lumen. In the lumen the hydronium ion reacts with bicarbonate ion, which is converted to resorbable nonionic CO_2.

There is also active transport of organic cations and anions into the lumen *(tubular secretion),* each by a separate system. These active transport systems are extremely important in the excretion of a number of drugs; for example, penicillin G is secreted rapidly by the anion transport system, and tetraethylammonium ion by the cation transport system. Probenecid is an inhibitor of anion secretion and, hence, decreases the rate of loss of penicillin from the body.

As the filtrate travels through the thin segment it becomes concentrated, especially at the bottom, as a result of active resorption and a countercurrent-distribution effect enabled by the recurrent and parallel arrangement of the ascending segment, the parallel orientation of the collecting duct, and the similar recurrent geometry of the associated capillaries.

In the thick segment of the ascending loop of Henle, both sodium and chloride are transported actively.

In the distal tubule, sodium resorption occurs partly in *exchange* for potassium *(potassium secretion)* and for hydronium ions. Adrenal mineralocorticoids promote distal tubular sodium resorption and potassium and hydronium secretion. *Ammonia secretion* also occurs, so that the urine either may be acidified or alkalinized, according to acid-base and electrolyte requirements.

Water is resorbed selectively from the distal end of the distal convoluted tubule and the collecting ducts; water resorption is under the control of the antidiuretic hormone.

Drugs also may be resorbed in the distal tubule; the pH of the urine there is extremely important in determining the rate

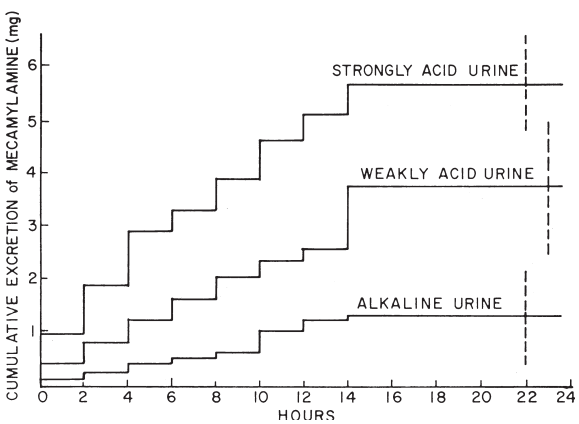


Figure 57-18. The effect of urinary pH on the mean cumulative excretion in man of mecamylamine during the first day after oral administration of 10 mg. *Vertical broken lines:* standard deviation. (From Milne MD, et al. *Clin Sci* 1957; 16:599.)

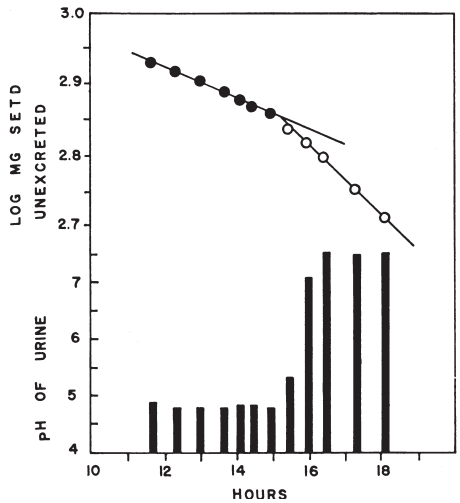


Figure 57-19. The effect of urinary pH on the excretion of sulfaethidole in a human subject after oral administration of 2 g. *Bars* (lower half): urinary pH; *circles* (open and closed, top): log of the amount of drug remaining in the body; *negative slopes* (of lines defined by the circles): a function of the rate constant of excretion. Note the abrupt increase in rate when the urinary pH is changed from acidic to neutral or slightly alkaline. (From Kostenbauder HB, et al. *J Pharm Sci* 1962; 51:1084.)

of resorption, in accordance with the principle of non-ionic diffusion and pH partition. The pH of the tubular fluid also affects the tubular secretion of drugs.

As an example of the importance of urinary pH, in humans the secondary amine mecamylamine is excreted more than four times faster when the urinary pH is below 5.5 than when it is above 7.5; Figure 57-18 illustrates the effect of urinary pH on the excretion of this amine. The effect of urinary pH on the excretion of a weak acid, sulfaethidole, is shown in Figure 57-19.

The urinary pH and, hence, drug excretion may fluctuate widely according to the diet, exercise, drugs, time of day, and other factors. Obviously, the excretion of weak acids and bases can be controlled partly with acidifying or alkalinizing salts, such as ammonium chloride or sodium bicarbonate, respectively. Comparative studies on potency and efficacy in man have demonstrated the importance of controlling urinary pH. Urinary pH is important only when the drug in question is a weak acid or base of which a significant fraction is excreted. The plasma levels will change inversely to the excretory rate. For example, it has been shown clinically with quinidine that alkalinization of the urine not only decreases the urine concentration but also increases the plasma concentration and toxicity.

The collecting duct also resorbs sodium and water, secretes potassium, and acidifies and concentrates the urine. Antidiuretic hormone (ADH) controls the permeability to water of both the collecting duct and the distal tubule.

Renal clearance and the kinetics of renal elimination are discussed in Chapter 58.

BILIARY EXCRETION AND FECAL ELIMINATION— Many drugs are secreted into the bile and then pass into the intestine. A drug that is passed into the intestine via the bile may be reabsorbed and not lost from the body. A drug conjugate entering the intestine may be deconjugated by enzymes and the parent drug reabsorbed. This cycle of biliary secretion and intestinal resorption is called *enterohepatic circulation.* Examples of drugs enterohepatically circulated are morphine, and the penicillins. The biliary secretory systems greatly resemble those of the kidney tubules. The enterohepatic system may provide a considerable reservoir for a drug.

If a drug is not absorbed completely from the intestine, the unabsorbed fraction will be eliminated in the feces. An unabsorbable drug that is secreted into the bile will likewise be eliminated in the feces. Such fecal elimination is called *fecal*

excretion. Only rarely are drugs secreted into the intestine through the succus entericus (intestinal secretions), although a number of amines are secreted into gastric juice.

ALVEOLAR EXCRETION—The large alveolar area and high blood flow make the lungs ideal for the excretion of appropriate substances. Only volatile liquids or gases are eliminated from the lungs. Gaseous and volatile anesthetics essentially are eliminated completely by this route. Only a small amount of ethanol is eliminated by the lungs, but the concentration in the alveolar air is related so constantly to the blood alcohol concentration that the analysis of expired air is acceptable for legal purposes. The high aqueous solubility and relatively low vapor pressure of ethanol at body temperature account for the retention of most of the substance in the blood. Carbon dioxide from those drugs that are partly degraded also is excreted in the lungs.

PHARMACOKINETICS

Pharmacokinetics is the science that treats the rate of absorption, extent of absorption, rates of distribution among body compartments, rate of elimination, and related phenomena. Because of its importance, Chapters 58 and 59, *Basic Pharmacokinetics* and *Clinical Pharmacokinetics,* have been devoted to the subject.

DRUG INTERACTION AND COMBINATION

Frequently a patient may receive more than one drug concurrently. Case records show that surgical patients commonly receive more than 10 drugs, and the patient is often under the influence of several drugs at once. Multiple-drug administration also is common for patients hospitalized for infections and other disorders. Furthermore, a patient may be suffering from more than one unrelated disorder that demands simultaneous treatment with two or more drugs. In such instances, interactions are unsolicited and often unexpected.

In addition to the administration of drugs concurrently for their independent and unrelated effects, drugs are sometimes administered concurrently deliberately to make use of expected interactions.

TYPES OF INTERACTION AND REASONS FOR COMBINATION THERAPY

A drug may affect the response to another drug in a quantitative way. On one hand, the intensity of either the therapeutic effect, or side effect, may be augmented or suppressed. On the other hand, a qualitatively different effect may be elicited. The mechanisms of such interactions are many and are not always well understood. A drug may not necessarily affect either the quality or initial intensity or effect of another drug, but may cause significant to profound changes in the duration of action. The nature of this type of interaction generally is understood fairly well, although it may not yet have been ascertained for any particular drug combination. The deliberate use of combined interacting drugs is most valid when the mechanism of the interaction is understood and the combined effects are both quantifiable and predictable. The rationales of drug combination and the principles involved are discussed below.

COMBINATIONS TO INCREASE INTENSITY OF RESPONSE OR EFFICACY—Sometimes the basis for the action of one drug to increase the intensity of response to another is well understood, but often the reason for a positive interaction is obscure. A terminology has arisen that frequently is not only enlightening as to mechanisms and principles but which also is somewhat confusing.

Drugs that elicit the same quality of effect and are mutually interactive are called *homergic,* regardless of whether there is anything in common between the separate response systems. Thus, the looseness of the term admits a pressor response consequent to an increase in cardiac output to be homergic with one resulting from arteriolar constriction, even though there is not one common responsive element, the blood pressure itself being but a passive indicator. However, homergic drugs usually have in common at least part of a response system. Thus, both norepinephrine and vasopressin stimulate some of the same vascular smooth muscle, even though they do not excite the same receptors.

Two homergic drugs can be agonists of the same receptor, so that the entire response system is common to both. Such drugs are called *homodynamic*. As discussed under *Drug Receptors and Receptor Theory*, homodynamic drugs will generate dose–intensity of effect curves with parallel slopes but not necessarily with identical maxima or efficacies, if one of the drugs is a partial agonist.

From mass-law kinetics and dose-effect data of the separate drugs, it is possible to predict the combined effects of two agonists to the same receptor. If both drugs are full agonists, theory predicts that an *EDx* of Drug *A* added to an *EDy* of Drug *B* should elicit the same effect as an *EDy* of Drug *A* added to an *EDx* of Drug *B*. An example is shown in Figure 57-20. Dose-percentage data with homodynamic drugs can be treated in the same way.[21]

Drugs whose combined effects fit the above conditions are called *additive*. If the response to the combination exceeds the expected value for additivity, the drugs are considered to be *supra-additive*. Purely homodynamic drugs do not show supra-additivity; however, if one drug in the pair has an additional action to affect the concentration or penetration of the other or to prime the response system in some way, two agonists to the same receptor may exhibit supra-additivity. Two homergic drugs are *infra-additive* if their combined effect is less than expected from additivity. As with supra-additivity, infra-additivity must involve an action elsewhere than on a common receptor.

Two drugs are said to be *summative* if a dose of drug that elicits response x added to a dose of another drug that elicits response y gives the combined response $x + y$. Very little significance usually can be attached to summation. Unless the dose-intensity curve of each drug is linear, rather than log-linear,

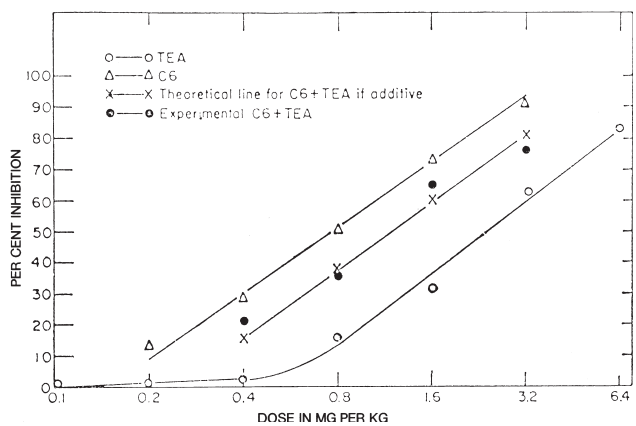


Figure 57-20. Additive inhibitory effects of tetraethylammonium (TEA) and hexamethonium (C6) on the superior cervical ganglion of the cat. The theoretical line for additivity was calculated on the basis that an increment of TEA added to an *EDx* of C6 should have the same effect as if it were added to an *EDx* of TEA. When TEA and C6 were administered together, an equal amount of each was given. The dose is the sum of the doses of the two components. (From Harvey SC. *Arch Intern Pharmacodyn* 1958; 114:232.)

summation cannot be predicted from the two curves. When summation does occur with the usual clinical doses of two drugs, it almost never occurs over the entire dose range; indeed, if the dose of each of the two drugs is greater than an ED50, summation is theoretically impossible unless it is possible to increase the maximal response. At best, summation is an infrequent clinical finding, limited to one or two doses.

Two drugs are said to be *heterergic* if the drugs do not cause responses of the same quality. When heterergy is positive, ie, the response to one drug is enhanced by the other, *synergism* is said to occur. The word often has been used to describe any positive interaction, but it should be used only to describe a positive interaction between heterergic drugs. The term *potentiation* has been used synonymously with synergism, but misuse of the term has led to the recommendation that the term be dropped. Synergism is often the result of an effect to interfere with the elimination of a drug and, thus, to increase the concentration; synergism also may result from an effect on penetration or on the responsivity of the effector system. Examples of a synergistic effect, in which responsivity is enhanced, are the action of adrenalcorticoids to enhance the vasoconstrictor response to epinephrine and the increase of epinephrine- induced hyperglycemia consequent to impairment by theophylline of the enzymatic destruction of the cAMP that mediates the response.

In clinical practice two homodynamic drugs rarely are coadministered for the purpose of increasing the response, since a sufficient dose of either drug should be able to achieve the same effect as a combination of the two. Most clinical combinations with positively interacting drugs involve heterergic drugs.

COMBINATIONS TO DECREASE INDIVIDUAL DOSES AND TOXICITY—When homodynamic drugs are coadministered, it is usually for the purpose of decreasing toxicity. If the toxicities of two homodynamic drugs are infra-additive, the toxicity of combined partial doses of the two drugs often will be less than with full doses of either drug. This principle is valid for trisulfapyrimidines mixture (see RPS-18, page 1181).

COMBINATIONS TO ATTACK A DISEASE COMPLEX AT DIFFERENT POINTS—With many diseases, more than one organ or tissue may be affected or events at more than one locus may bear upon the ultimate perturbation. For example, in duodenal ulcer, psychic factors appear to increase activity in the vagus nerve, which modulates gastric secretion, so that it is rational to explore the effects of sedatives, ganglionic blocking drugs, antimuscarinic drugs, and antacids, singly and in combination. In heart failure the decrement in renal plasma flow and changes in aldosterone levels promote the retention of salt and water, so that diuretics and digitalis usually are employed concomitantly. Pain, anxiety, and agitation or depression are frequent accompaniments of various pathological processes, so that it is to be expected that analgesics, tranquilizers, sedatives, or antidepressives frequently will be given at the same time, along with other drugs intended to correct the specific pathology.

COMBINATIONS TO ANTAGONIZE UNTOWARD ACTIONS—The side effects of a number of drugs can be prevented or suppressed by other drugs. An antagonist may compete with the drug at the receptor that initiates the side effect, depress the side-effector system at a point other than the receptor, or stimulate an opposing system.

Antagonism at the receptor is *competitive antagonism* if the antagonist attaches at the same receptor group as the agonist (see page 1104). Antagonism at a different receptor group or inhibition elsewhere in the response system is *noncompetitive antagonism*. Both competitive and noncompetitive antagonism are classified as *pharmacological antagonism*. The stimulation of an opposing system is *physiological antagonism*.

Examples of pharmacological antagonism are the use of atropine to suppress the muscarinic effects of excess acetylcholine consequent to the use of neostigmine and the use of antihistaminics to prevent the effects of histamine liberated by tubocurarine. Examples of physiological antagonism are the use of amphetamine to correct partially the sedation caused by anticonvulsant doses of phenobarbital and the administration of ephedrine to correct hypotension resulting from spinal anesthesia.

COMBINATIONS THAT AFFECT ELIMINATION—Only a few drugs presently are used purposefully to elevate or prolong plasma levels by interfering with elimination, although continued interest in such drugs probably will increase the number.

Probenecid, which already has been mentioned to antagonize the renal secretion of penicillin, was introduced originally for this purpose. However, because penicillin G is inexpensive and available in repository forms as well as oral forms (obviating the need for injection), it is less imperative to retard the excretion of penicillin. The low, nonallergenic toxicity of penicillin permits very large doses to be given without concern for the high plasma concentrations that result, which also means that there is little necessity for increasing the biological half-life of the drug. Consequently, probenecid is not used routinely today in combination with penicillin.

The use of vasoconstrictors to increase the sojourn of local anesthetics at the site of infiltration continues, but few other clinical examples of the deliberate use of one drug to interfere with either the distribution or elimination of another can be cited. Nevertheless, the subject of the effect of one drug on the elimination of another has become immensely active. Innumerable drugs affect the fate of others, and the therapist must be aware of such interactions.

Drugs that induce cytochrome P450s and other drug- metabolizing enzymes enhance the elimination of drugs that are metabolized by the liver. There would be very little point ordinarily in soliciting combinations that would shorten the duration of action or lower plasma levels, unless it were to reduce an overdosage. However, since such combinations are used unwittingly or unavoidably, this type of interaction is of great clinical importance.

Drugs that inhibit cytochrome P450 will, of course, reduce the metabolism of a wide range of additional drugs and serve to prolong or elevate plasma concentration.

COMBINATIONS TO ALTER ABSORPTION—In the section *Vehicles and Absorption Adjuvants*, it was mentioned that certain substances facilitate the absorption of others. The use of such absorption adjuvants generally is included under the subject of formulation rather than under drug combination. Although drugs that increase blood flow, motility, etc, have an effect to increase the rate of absorption, the use of such drugs so far has not proved to be very practical. When it is desired to slow the absorption of drugs, various physical or physicochemical means prove to be more effective and less troublesome than drug combinations.

Fixed Combinations of Drugs

Concomitant treatment with two or more drugs frequently is unnecessary, and generally, it immeasurably complicates therapy and the evaluation of response and toxicity. Nevertheless, it is often warranted, even essential, and cannot be condemned categorically. However, with fixed-dose or fixed-ratio combinations, in which the drugs are together in the same preparation, there are certain disadvantages, except for a few rare instances such as trisulfapyrimidines.

The disadvantages are as follows: patients differ in their responsivity or sensitivity to drugs, and adjustments in dosage or dose-interval may be necessary. If adjustment of only one component of the mixture is required, it is undesirable that the schedule of the second component be adjusted obligatorily, as it is in a fixed combination. According to which way the dose is adjusted, either toxicity or loss of the therapeutic effect may result. Furthermore, when adverse effects to either component occur, both drugs must be discontinued. The fixed combination denies the physician flexible control of therapy. Especially when one component in a mixture is superfluous yet potentially toxic, as is often the case, the promotion of fixed combinations is reprehensible. However, the separate administration of drugs used in

combination often complicates treatment for patients, who, in an outpatient situation and sometimes in the hospital, may not take all of their medication or may take it at inappropriate intervals. The resulting consequences may be worse than those of fixed combinations in certain instances. Consequently, a summary dismissal of fixed combinations is unwarranted. Rather, the fundamentals of pharmacokinetics and clinical experience must be brought together with biopharmaceutics to analyze present combinations and to predict possible new allowable combinations.

DANGERS IN MULTIPLE-DRUG THERAPY

Some objections to fixed-dose combinations were stated above. Also the unanticipated effects of drug combinations have been touched upon, particularly with respect to effects upon elimination. But it should be made clear that more is at stake than simply the biological half-life of a drug. An example is given of the grave clinical consequences of the effect of phenobarbital enhancing the biotransformation of warfarin. Other examples of dangerous interactions, such as the effect of several antidepressants in greatly synergizing catecholamines, may be cited. Even some antibiotics antagonize each other and increase mortality.

In addition to the obvious pitfalls posed by the interactions themselves, the use of multiple-drug therapy fosters careless diagnosis and a false sense of security in the number of drugs employed. Multiple-drug therapy should never be employed without a convincing indication that each drug is beneficial beyond the possible detriments or without proof that a therapeutically equivocal combination is definitely harmless. Finally, the expense to the patient warrants consideration.

REFERENCES

1. Clark AJ. *J Physiol (London)* 1926; 61:547.
2. Ariens EJ, ed. *Molecular Pharmacology,* vol 1. New York: Academic, 1964, p 176.
3. Stephenson RP. *Br J Pharmacol* 1956; 11:379.
4. Rang HP. *Br J Pharmacol* 1973; 48:475.
5. Colquhoun D. In: Rang HP, ed. *Drug Receptors.* Baltimore: University Park, 1973.
6. Schanker LS. *Adv Drug Res* 1964; 1:71.
7. Brodie BB, et al. *J Pharmacol Exp Ther* 1960; 130:20.
8. Truitt EB, et al. *J Pharmacol Exp Ther* 1950; 100:309.
9. Lillehei JP. *JAMA* 1968; 205:531.
10. Martin AN, et al. *Physical Pharmacy,* 2nd ed. Philadelphia: Lea & Febiger, 1969, pp 247, 253.
11. Jacobs MH. *Cold Spring Harbor Symp Quant Biol* 1940; 8:30.
12. Schanker LS. *Pharmacol Rev* 1961; 14:501.
13. Brodie BB, Hogben CA. *J Pharm Pharmacol* 1957; 9:345.
14. Hogben CA. *Fed Proc* 1960; 19:864.
15. Albert A. *Pharmacol Rev* 1952; 4:136.
16. Ariens EJ, et al. In: *Molecular Pharmacology,* vol 1. Ariens EJ, ed. New York: Academic, 1964, p 7.
17. Nelson DR, et al. *DNA Cell Biol* 1993; 12:1.
18. Milne MD, et al. *Clin Sci* 1957; 16:599.
19. Kostenbauder HB, et al. *J Pharm Sci* 1962; 51:1084.
20. Harvey SC. *Arch Intern Pharmacodyn* 1958; 114:232.
21. Weaver LC, et al. *J Pharmacol Exp Ther* 1955; 113:359.

BIBLIOGRAPHY

Anders MW. *Bioactivation of Foreign Compounds.* New York: Academic, 1985.

Bend JR, Serabjit-Singh CJ, Philpot RM. The pulmonary uptake accumulation and metabolism of xenobiotics. *Annu Rev Pharmacol Toxicol* 1985; 25:97.

Benford D, et al. *Drug Metabolism: From Molecules to Man*. New York: Taylor & Francis, 1987.

Bertolino M, Llinas R. The central role of voltage activated and receptor operated calcium channels in neuronal cells. *Annu Rev Pharmacol Toxicol* 1992; 32:399.

Black JW, et al, eds. *Perspectives on Receptor Classification.* New York: Liss, 1987.

Boelsterli U. *Mechanistic Toxicology: The Molecular Basis of How Chemicals Disrupt Biological Targets.* New York: Taylor and Francis, 2003

Caldwell J, Jakoby WB. *Biological Basis of Detoxification*. New York: Academic, 1983.

Coulson CJ. *Mechanisms of Drug Action*. New York: Taylor & Francis, 1987.

Dean PM. *Molecular Foundations of Drug-Receptor Interaction*. Cambridge: Cambridge University Press, 1987.

Denison MS, Nagy SR. Activation of the aryl hydrocarbon receptor by structurally diverse exogenous and endogenous chemicals *Annu Rev Pharmacol Toxicol* 2003; 43:309.

Ding X, Kaminsky LS. Human extrahepatic cytochromes P450: function in xenobiotic metabolism and tissue selective chemical toxicity in the respiratory and gastrointestinal tract. *Annu Rev Pharmacol Toxicol* 2003; 43:149.

Finean JB, Michell RH, eds. *Membrane Structure.* Amsterdam: Elsevier/North Holland, 1981.

Gibson GG, Skett PL. *Introduction to Drug Metabolism,* 2nd ed. London: Chapman & Hall, 1994.

Gilman AG. G proteins: transducers of receptor generated signals. *Annu Rev Biochem* 1987; 56:615.

Lewis DFV. *Guides to Cytochromes P450.* New York: Taylor & Francis, 2002.

Gregoriadis G, Senior J. *Targeting of Drugs with Synthetic Systems*. New York: Plenum, 1986.

Hulme EC, Birdsall NJM, Buckley NJ. Muscarinic receptor subtypes. *Annu Rev Pharmacol Toxicol* 1990; 30:633.

Ioannides C, ed. *Cytochromes P450, Metabolic and Toxicological Aspects*. Boca Raton, FL: CRC Press, 1996.

Jakoby WB, et al, eds. *Metabolic Basis of Detoxification*. New York: Academic, 1982.

Kalow W. Pharmacogenetics in biological perspective. *Pharmacol Rev* 1997; 49:369

Karlin A. Anatomy of a receptor. *Neurosci Comment* 1983; 1:111.

Kenakin TP. The classification of drugs and drug receptors in isolated tissues. *Pharmacol Rev* 1984; 36:165.

Kenakin TP. *Pharmacological Analysis of Drug Receptor Interaction*. New York: Raven Press, 1987.

La Du B, et al. *Fundamentals of Drug Metabolism and Drug Disposition*. Baltimore: Williams & Wilkins, 1971.

Lamble JW, Abbott AC, eds. *Receptors Again!* Amsterdam: Elsevier, 1984.

Lefkowitz RJ, ed. *Receptor Regulation*. London: Chapman & Hall, 1981.

Levine RR. *Pharmacology: Drug Actions and Reactions,* 4th ed. Boston: Little, Brown, 1990.

Limbird LE. *Cell Surface Receptors: A Short Course on Theory and Methods.* Boston: Nijhoff, 1986.

Loh HH, Smith AP, Birnbammer L. Molecular characterization of opioid receptors G proteins in signal transduction. *Annu Rev Pharmacol Toxicol* 1990; 30:123.

Martonosi AN. *Membranes and Transport*. New York: Plenum, 1982.

Meyer UA. Drugs in special patient groups: clinical importance of genetics in drug effects. In: Melman KL, et al, eds. *Clinical Pharmacology,* 3rd ed. New York: McGraw-Hill, 1992.

Mulder GJ, ed. *Conjugation Reactions in Drug Metabolism.* New York: Taylor & Francis, 1990.

Mulder GJ: Glucuronidation and its role in regulation of biological activity of drugs. *Annu Rev Pharmacol Toxicol* 1992; 32:25.

Nguyen T, Sherratt PJ, Pickett CB. Regulatory mechanisms controlling gene expression mediated by the antioxidant response element *Annu Rev Pharmacol Toxicol* 2003; 43:233.

O'Dowd BF, et al. Structure of the adrenergic and related receptors. *Annu Rev Neurosci* 1989; 12:67.

Olson RW, Venter JC, eds. *Benzodiazepine/GABA Receptors and Chloride Channels.* New York: Liss, 1986.

Ortiz de Montellano PR, ed. *Cytochrome P450. Structure, Mechanism, and Biochemistry,* 3rd ed. New York: Kluwer Academic Plenum, 2003.

Parkinson A. Biotransformation of xenobiotics. In: *Casarett and Doull's Toxicology: The Basic Science of Poisons,* 6th ed. New York: McGraw Hill, 2001.

Post G, Crooke ST, eds. *Mechanisms of Receptor Regulation.* New York: Plenum, 1986.

Pratt WB, Taylor P. *Principles of Drug Action*. New York: Churchill Livingstone, 1990.

Putney JW Jr, ed. *Phosphoinositides and Receptor Mechanisms.* New York: Liss, 1986.

Roche EB, ed. *Bioreversible Carriers in Drug Design.* New York: Pergamon, 1987.

Roth SH, Miller KW, eds. *Molecular and Cellular Mechanisms of Anesthetics.* New York: Plenum, 1986.

Sandler M, ed. *Enzyme Inhibitors As Drugs*. Baltimore: University Park, 1980.
Schmucker DL. Aging and drug disposition. *Pharmacol Rev* 1985; 37:133.
Schou JS, et al, eds. *Drug Receptors and Dynamic Processes in Cells*. New York: Raven, 1986.
Stein WD. *Transport and Diffusion Across Cell Membranes*. Orlando: Academic, 1986.
Stoughton RB. Percutaneous absorption of drugs. *Annu Rev Pharmacol Toxicol* 1989; 29:55.
Stroud RM. Acetylcholine receptor structure. *Neurosci Comment* 1983; 1:124.
Sueyoshi T, Negishi M. Phenobarbital response elements of cytochrome P450 genes and nuclear receptors. *Annu Rev Pharmacol Toxicol* 2001; 41:123.
Testa, B, ed. *Advances in Drug Research,* vols 14, 15. London: Academic, 1985, 1986.
Thummel KE, Wilkinson GR. In vitro and in vivo drug interactions involving human CYP3A. *Annu Rev Pharmacol Toxicol* 1998; 38:389.
Triggle DJ, Janis RA. Calcium channel ligands. *Annu Rev Pharmacol Toxicol* 1987; 27:347.
Tukey RH, Strassburg CP. Human UDP-glucuronosyltransferases: metabolism, expression and disease. *Annu Rev Pharmacol Toxicol* 2000; 40:581.
Venter JC, Harrison LC, eds. *Molecular and Chemical Characterization of Membrane Receptors*. New York: Liss, 1984.
Wardle EN. *Cell Surface Science in Medicine and Pathology*. New York: Elsevier, 1985.
Yamazaki M, Suzuki H, Sugiyama Y. Recent advances in carrier mediated hepatic uptake and biliary excretion of xenobiotics. *Pharmacut Res* 1996; 13:497.
Zaki Y, et al. Opioid receptor types and subtypes: the δ receptor as a model. *Annu Rev Pharmacol Toxicol* 1996; 36:379.

CHAPTER 58

Basic Pharmacokinetics and Pharmacodynamics

Raymond E Galinsky, PharmD
Craig K Svensson, PharmD, PhD

The goal of pharmacotherapy is to provide optimal drug therapy in the treatment or prevention of disease. A major barrier to the achievement of this goal is the large variability in pharmacological effect that is observed following drug administration (Fig 58-1).[1] The ability to implement drug therapy in a safe and rational manner necessitates an understanding of the factors that cause this variability. One of the most important factors is the concentration of drug that is achieved at the site of action.

THE CRITICAL NATURE OF THE CONCENTRATION VERSUS EFFECT RELATIONSHIP

The quantitative response to a drug depends highly on the concentration of drug at the site of action. In most situations, one cannot quantify drug concentration at the actual site of action. Rather, drug concentrations are measured in an easily accessible site that is believed to be in equilibrium with the site of action (eg, blood or one of its components). Figure 58-2[2] provides a good illustration of a drug whose pharmacological effect is particularly sensitive to changes in blood concentration. Numerous studies have been published that substantiate the critical nature of the concentration-effect relationship for a wide variety of drugs.

It is recognized now that drug therapy may be optimized by designing regimens that account for the concentration of a drug necessary to achieve a desired pharmacological response. However, there is often significant difficulty in achieving such target concentrations. In particular, it often is observed that if a fixed dose of a drug is administered to a group of individuals, the drug concentration measured in plasma can vary widely. For example, the peak concentration of 6-mercaptopurine achieved in a group of 20 patients who received a standard 1 mg/m^2 dose is shown in Figure 58-3.[3] The concentrations ranged from 0 to 660 ng/mL. Taken together, this suggests that variability in drug concentration is a major source of variability in drug effect, and there may be a significant degree of variability among individuals in the drug concentrations produced by a given dose of drug.

A basic understanding of the factors that control drug concentration at the site of action is important for the optimal use of drugs. This is the area of study referred to as *pharmacokinetics,* which is the study of the time course of drug absorption, distribution, metabolism, and elimination.

DRUG CONCENTRATION VERSUS TIME PROFILE

Blood (or its components, plasma or serum) represents the most frequently sampled fluid used to characterize the pharmacokinetics of drugs. Drug concentration in the blood is the sum of several processes (Fig 58-4).[4] Initially, visual characterization of the processes controlling the concentration of drug in the blood can be made by constructing a drug concentration versus time profile (ie, a plot of drug concentration in the blood versus time). As can be seen from Figure 58-5, several useful pieces of information can be derived from such a profile. For example, the time at which the peak concentration occurs can be approximated and the peak concentration quantified. If the minimum concentration needed to maintain a desired effect is known, the onset and duration of effect also can be approximated. While useful information can be drawn casually from a simple graph as depicted in Figure 58-5, a more rigorous description of the pharmacokinetics of a drug is necessary to achieve the accuracy in dosage regimen design required for the safe and effective use of drugs. This higher degree of accuracy necessitates the development of mathematical models for describing the time course of absorption, distribution, metabolism, and elimination.

PHARMACOKINETIC MODELS

One of the primary objectives of pharmacokinetic models is to develop a quantitative method to describe the relationship of drug concentration or amount in the body as a function of time. The complexity of the pharmacokinetic model will vary with the route of administration, the extent and duration of distribution into various body fluids and tissues, the processes of elimination, and the intended application of the pharmacokinetic model. Often, numerous potential mathematical models exist for a particular drug. In such cases, the simplest model that will adequately and accurately describe the pharmacokinetics of the drug is the model that should be chosen.

There are a wide variety of potential uses for pharmacokinetic models, which include

1. Prediction of drug concentration in blood/plasma or tissue.
2. Calculation of a dosage regimen.
3. Quantitative assessment of the effect of disease on drug disposition.
4. Elucidation of the mechanism of disease-induced alterations in drug disposition.
5. Determination of the mechanism for drug-drug interactions.
6. Prediction of drug concentration versus effect relationships.

There are three primary types of pharmacokinetic models: compartmental, noncompartmental, and physiological.

Compartmental models describe the pharmacokinetics of drug disposition by grouping body tissues that are kinetically indistinguishable and describe the transfer of drug between body tissues in terms of rate constants.

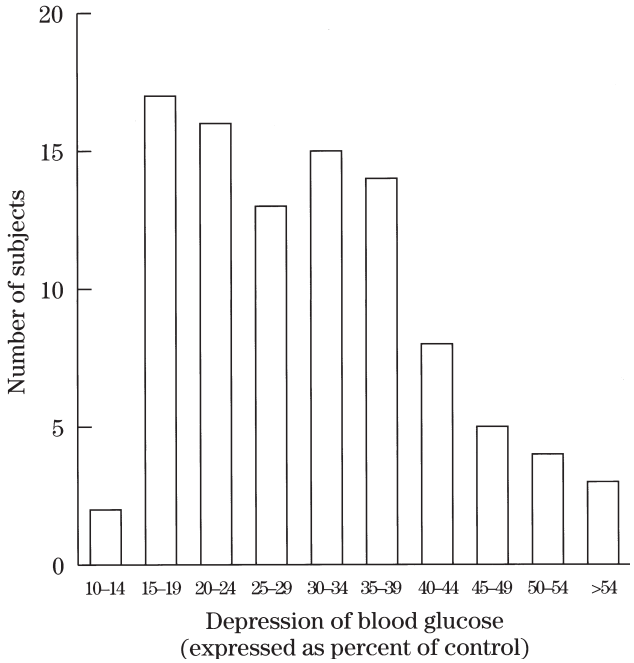


Figure 58-1. Decrease in blood glucose in 97 subjects 30 min after an intravenous dose of 1 g of tolbutamide. Note the large variability observed after the equivalent dose was administered in this group. (Data from Swerdloff RS, et al. *Diabetes* 1967; 16:161.)

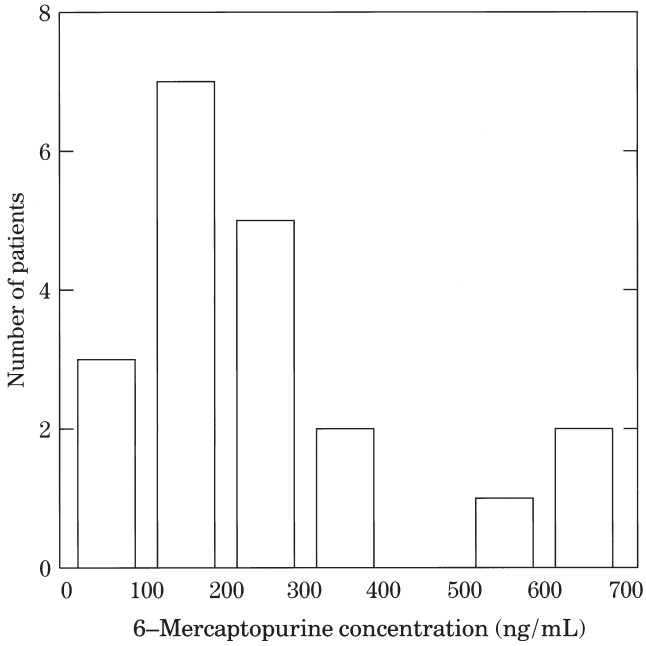


Figure 58-3. Distribution of peak 6-mercaptopurine concentrations achieved in a group of 20 patients receiving an oral dose of 1 mg/m^2. (Data from Sulh H, et al. *Clin Pharmacol Ther* 1986; 40:604.)

Noncompartmental models describe the pharmacokinetics of drug disposition using time- and concentration-averaged parameters.

Physiological models attempt to describe drug disposition in terms of realistic physiological parameters, such as blood flow and tissue-partition coefficients.

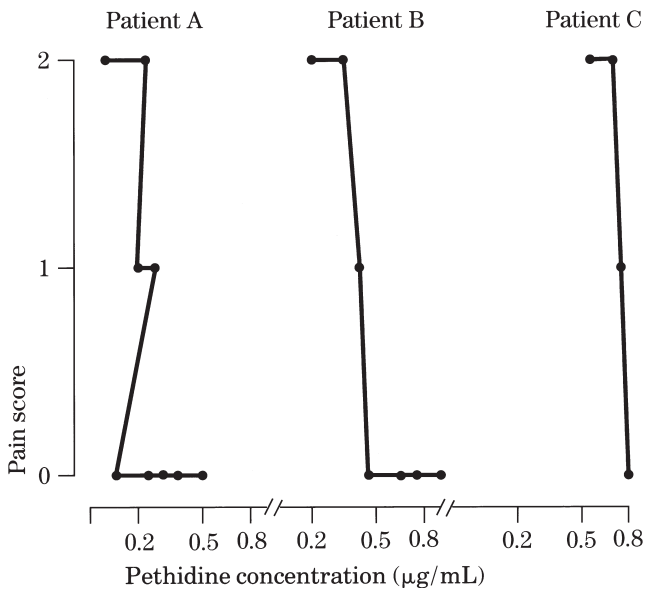


Figure 58-2. Blood-pethidine concentration-response curves for three individual patients, illustrating a typical range in interpatient responses. (From Edwards DJ, et al. *Clin Pharmacokinet* 1982; 7:421.)

RATES AND ORDERS OF REACTIONS

Many pharmacokinetic models use parameters that are analogous to rate constants in chemical kinetics. For example, consider the case of a drug (D) that is metabolized to a metabolite (M).

$$D \rightarrow M$$

This reaction may be described as a function of either the disappearance of the drug or as a function of the appearance of the metabolite. If the *amount* of the drug that is converted to a metabolite is a constant with respect to time, the reaction is said to be *zero-order* and is expressed as

$$\frac{-dD}{dt} = k_0 \qquad (1)$$

where K_0 is the zero-order rate constant with units of mass per time (eg, mg/min). A plot of the time-course of the amount of the drug in the body that is converted to a metabolite by zero-order kinetics is shown in Figure 58-6. Integration of Equation 1 yields an equation for a straight line, which describes the amount of the drug in the body at any time (t):

$$\text{Amount}_t = -k_0 t + \text{Amount}_{t=0} = -k_0 t + \text{Dose} \qquad (2)$$

Zero-order rate processes typically are found when an enzyme or transport system becomes saturated and the rate process becomes constant and cannot be increased by increases in the concentration of substrate. Zero-order rate processes are typical of constant-rate intravenous infusions and prolonged- release dosage forms.

If the amount of the drug in the body is converted to a metabolite at a rate that is a constant *fraction* of the amount of the drug in the body, the conversion of D to M is said to be a first-order reaction described by

$$\frac{dD}{dt} = -kD \qquad (3)$$

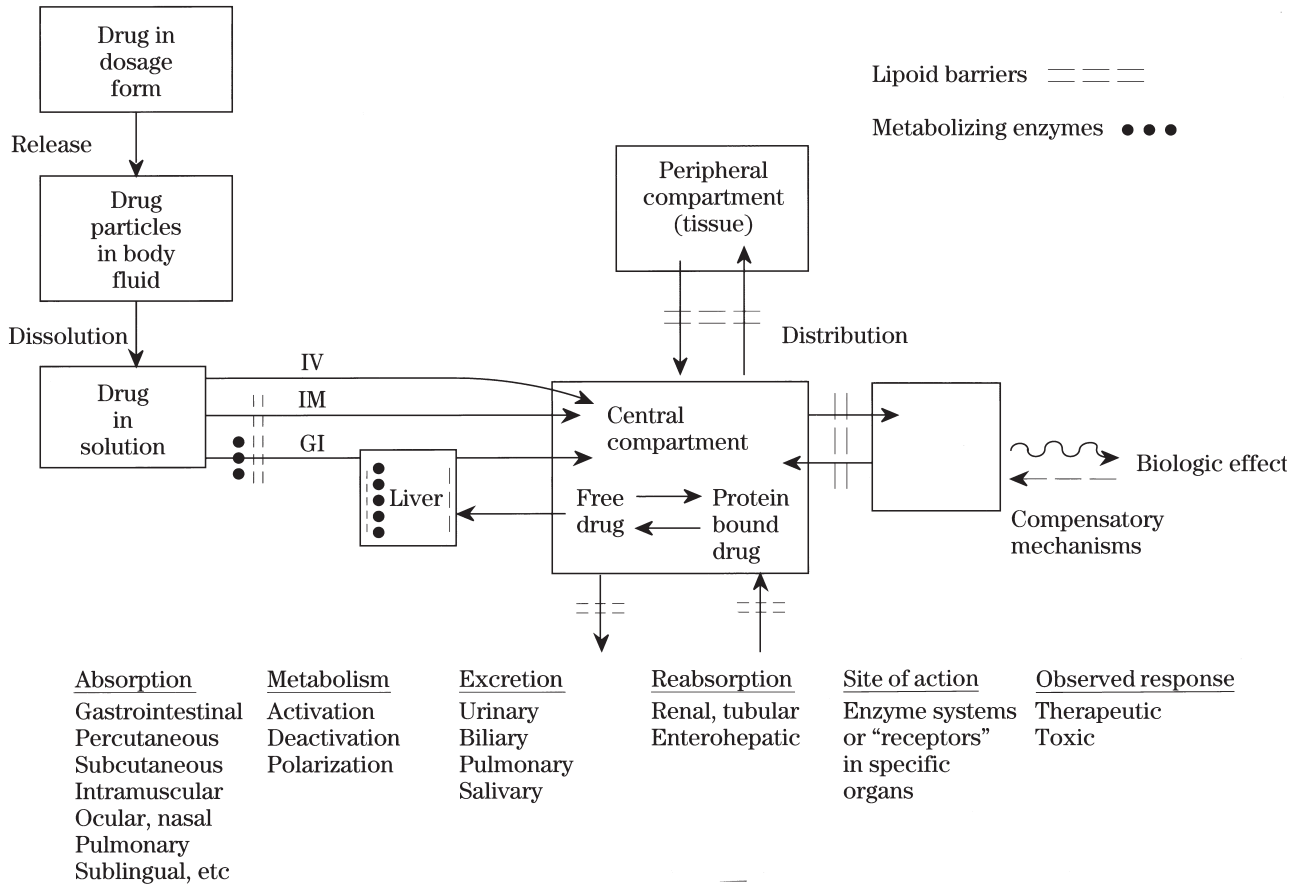


Figure 58-4. Diagram illustrating the factors that influence onset, duration, and intensity of drug effects. Note that the drug must dissolve before being absorbed and that it passes across many lipoid barriers and some metabolizing systems before reaching the site of action. (From Barr WH. *Am J Pharm Educ* 1968; 52:958.)

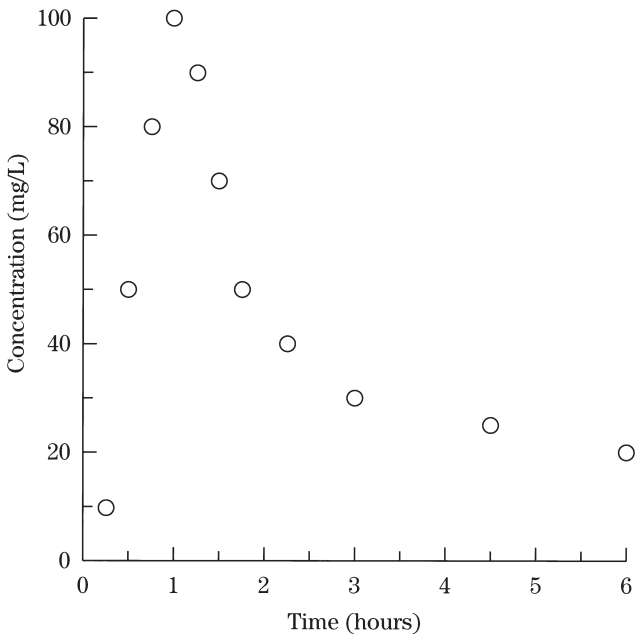


Figure 58-5. Hypothetical plot of drug-concentration data after oral administration of a drug.

where k is the first-order rate constant expressed in units of reciprocal time (eg, min^{-1}). Rearrangement of Equation 3 leads to

$$\frac{dD}{D} = -kdt \tag{4}$$

and integration of this expression yields

$$\int_0^t \frac{dD}{D} \Rightarrow \ln D = -kt + \ln D_0 \tag{5}$$

where ln is the natural logarithm. This equation also can be expressed in the exponential form

$$D_t = D_0 e^{-kt} \tag{6}$$

Graphically, the integrated form usually is expressed in terms of $\log_{10}$ rather than in natural logarithms (see Fig 58-6):

$$\log D = \frac{-kt}{2.303} + \log D_0 \tag{7}$$

ANALYTICAL CONSIDERATIONS

Any discussion of pharmacokinetics presumes that the drug concentrations can be determined with a high degree of accuracy and precision. One of the most frequent causes of high variability in pharmacokinetic parameters is poor data resulting

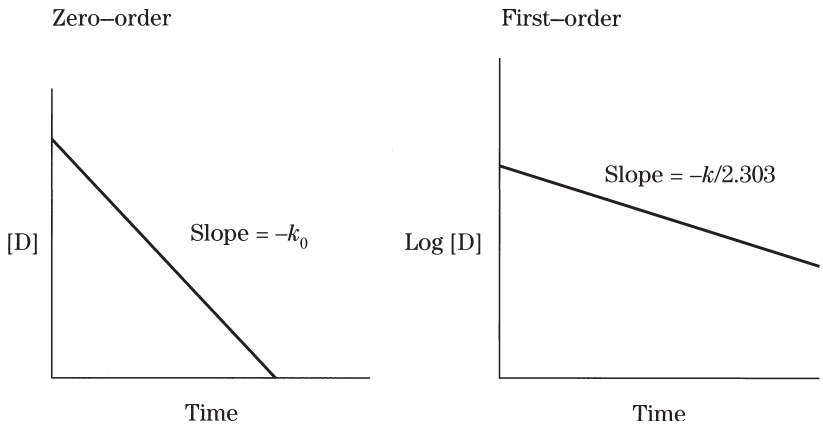


Figure 58-6. Plots illustrating a zero-order and a first-order reaction.

from imprecise analytical procedures. Evaluation of pharmacokinetic data in the literature must begin with an assessment of the validity of the assay used under the conditions in which the study was conducted. An assay must be tested for specificity, sensitivity, reproducibility, stability, and accuracy. Because drug metabolites are frequently present in the fluid to be measured and are similar in structure to the parent compound, differentiation of drug from any putative metabolites must be ensured.

INSTANTANEOUS INPUT WITH INSTANTANEOUS DISTRIBUTION

The disposition of a drug from its site of administration and its distribution and elimination from the body occurs via the vascular system. Most drugs are low-molecular-weight compounds of sufficient lipophilicity that they are able to distribute readily into the intra- and extracellular fluid compartments in the body. The transfer of drug from the circulation to these fluid compartments and then into tissues is called distribution. The pharmacokinetic parameter *volume of distribution* is a proportionality constant that relates drug concentration in a reference fluid, typically plasma, to the amount of drug distributed throughout the body.

$$\text{Volume of distribution } (V_D) = \frac{\text{Amount of drug in body } (D_B)}{\text{Drug concentration } (Cp)} \quad (8)$$

Drugs that distribute widely to tissues will have large volumes of distribution and low plasma concentrations relative to the dose administered, whereas drugs that are highly bound to plasma proteins (eg, warfarin, phenylbutazone) or do not readily enter cells (eg, amikacin) will have low volumes of distribution and high plasma concentrations relative to the administered dose.

Øie and Tozer[5] have developed a physiological model for expression of the apparent volume of distribution, which takes into account the extracellular water, including plasma- and protein-binding of the drug in both plasma and tissue. For an average 70-kg male, total body water is about 42 L, of which 3 L is plasma and 12 L is extracellular fluid space. Moreover, 55–60% of the albumin in the extracellular space is found outside of plasma. Thus, for drugs that are largely bound to albumin, the apparent volume of distribution can be expressed as

$$V_D = 7 + 8 f_u + V_T \left[\frac{f_u}{f_{u_T}}\right] \quad (9)$$

where f_u is the fraction of the drug in plasma that is unbound (often referred to as the *free fraction*), f_{uT} is the free fraction of drug in tissue, and V_T is the volume of intracellular tissue water. Equation 9 can be simplified further to

$$V_D = 7 + 8 f_u + 27 \left[\frac{f_u}{f_{u_T}}\right] \quad (10)$$

This model has been extremely useful in predicting the magnitude of changes in the apparent volume of distribution due to alterations in (1) plasma protein binding, (2) tissue protein binding, and (3) the volume of extracellular and intracellular fluid. For example, if a drug distributes into extracellular fluid but not intracellular fluid, the apparent volume of distribution can be expressed as

$$V_D = 7 + 8 f_u \quad (11)$$

and will vary between 7 and 15 L, depending upon the extent of plasma protein binding to albumin. For such a compound, with a relatively small volume of distribution, alterations in the plasma protein binding will not produce proportional changes in the apparent volume of distribution. Indeed, as reported by Williams et al,[6] the free fraction of tolbutamide in plasma increases in patients with cirrhosis by 28%, from 0.068 to 0.087, yet the apparent volume of distribution increases less than 10%, from 0.15 to 0.164 L/kg. Conversely, drugs with a volume of distribution greater than total body water indicate drug distribution and binding to tissue proteins and other cellular components. Such compounds also may be bound highly to plasma proteins. With drugs having a large volume of distribution (>50–100 L), the contribution of plasma and extracellular water space can be ignored, and Equation 3 simplifies to

$$V_D = 27 \left[\frac{f_u}{f_{u_T}}\right] \quad (12)$$

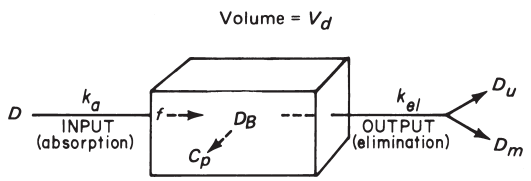


Figure 58-7. The open one-compartment pharmacokinetic model. The amount of the drug dose *(D)* that enters the body is D_B; in the case of intravenous injection, this is the entire dose, whereas in the case of extravascular administration, some fraction of the dose *(F)* is absorbed with a rate constant of k_a. The compartment has an apparent volume of distribution *(V_d)* into which drug distributes instantaneously to achieve a concentration of C_p. Drug is eliminated from the compartment with a rate constant k_{el}. D_u is the amount excreted into urine, feces, bile, expired air, sweat, milk, etc; D_m is the amount of drug metabolized.

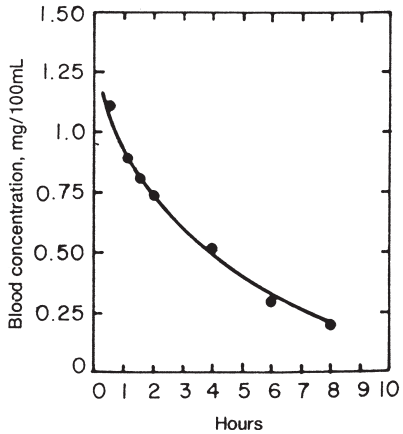


Figure 58-8. Elimination curve of average blood levels of theophylline in 11 human subjects after intravenous administration of 0.5 g aminophylline per 70 kg to each. (Data from Truitt EB Jr, et al. *J Pharmacol Exp Ther* 1950; 100:309.)

Changes in the plasma or plasma-free fraction will produce proportional changes in the apparent volume of distribution. For example, a twofold increase in the free fraction of drug in tissue, f_{uT}, will decrease the apparent volume of distribution by twofold. Less drug will be distributed to tissue, reflected by an increase in plasma concentrations. The volume necessary to account for the total amount of drug in the body will appear to have been decreased.

Once a drug is in the vascular system, it is transported by the blood to tissues where it can be eliminated from the circulation by distribution into tissue, metabolism by the tissue, or excretion from the tissue (see Fig 58-4). All of these processes lower the plasma concentration of drug. Each separate process may be described by a first-order rate constant, and the overall change in the plasma concentration is the net effect of all of these parallel, competing, first-order processes.

Intravenous injection of a drug that has nearly instantaneous distribution and first-order elimination can be described by an open one-compartment model (Fig 58-7). The body behaves as if it were a homogeneous compartment. In the one-compartment model, distribution is very rapid and can be considered instantaneous and is, therefore, ignored. After intravenous administration the plasma concentration declines exponentially according to

$$C = C_0 e^{-\lambda t} \tag{13}$$

where C_0 is the initial concentration and λ is the overall elimination rate constant. Such an exponential elimination of theophylline given intravenously, is shown in Figure 58-8.[7] According to Equation 13, if the data of Figure 58-8 are plotted on semilog paper, a straight line should result, and such a plot is shown in Figure 58-9. Several derived data can be obtained from a plot of log concentration versus time. Extrapolation to zero time (ie, the y-intercept) gives an estimated theoretical concentration in plasma at time zero, from which the apparent volume of distribution (V_D) can be estimated by simply dividing the dose by C_0. It is a theoretical concentration because neither the injection nor distribution are actually instantaneous.

The *half-life* of a drug is the time required to reduce the amount of drug in the body or the plasma concentration by 50%. For a first-order process, the half-life is constant and is independent of the starting value of the amount of drug in the body (or plasma concentration). The plasma half-life, $t_{1/2}$, can be determined directly from the graph or from the elimination rate constant, λ or k_{el}, by means of the relationship

$$t_{1/2} = \frac{0.693}{\lambda} = \frac{0.693}{k_{el}} \tag{14}$$

As shown, the half-life is related inversely to the elimination rate constant. When the elimination rate constant, k_{el}, is determined from the slope of the concentration versus time plot, one must keep in mind that the data need to be plotted on a semilog scale. From Figure 58-9, the k_{el} is determined to be 0.22 hr^{-1}. This is the instantaneous rate constant and indicates that 22% of the theophylline in the body is lost per hour. The rate constant for elimination (see Fig 58-7) is shown without reference to the route of elimination. It must be recognized that k_{el} represents the overall elimination by all competing, parallel pathways and is equal to the sum of the rate constants that define the various simultaneous (ie, parallel) contributory processes (eg, metabolism, renal excretion, or biliary secretion). Thus, the overall rate constant, $k_{el} = k_1 + k_2 + k_3 + \ldots k_N$, where $k_1 + k_2 + k_3 + \ldots k_N$ are the rate constants of the separate contributory processes.

Half-life is a clinically useful pharmacokinetic parameter in that it indicates when the next dose of a drug needs to be administered and is therefore helpful in designing an optimal dosing regimen. The half-life also is useful in determining

1. The fluctuation of plasma concentrations between doses;
2. The time required to reach steady-state equilibrium after beginning continuous drug administration; and
3. The persistence of drug in the system once drug administration has ceased.

Under some conditions, it is not possible to obtain plasma concentration data over sufficient time to obtain accurate estimates of the half-life for designing dosage regimens. The elimination rate constant, and hence the half-life, may be estimated from the excretion rate of unchanged drug. Because the first-order elimination rate constant is independent of the amount of drug in the body, the instantaneous excretion rate, dD_u/dt is directly proportional to the total amount of drug in the body.

$$\frac{dD_u}{dt} = k_e D_B \tag{15}$$

where D_B and D_u are the amount of drug in the body at time zero and the amount of drug excreted in the urine, respectively, and k_e is the urinary excretion rate constant. A plot of log dD_u/dt versus time yields a straight line with slope of $-k_e/2.3$. One also may estimate the half-life from urinary excretion data using the cumulative amount of drug excreted (sigma-minus) method. Using this approach

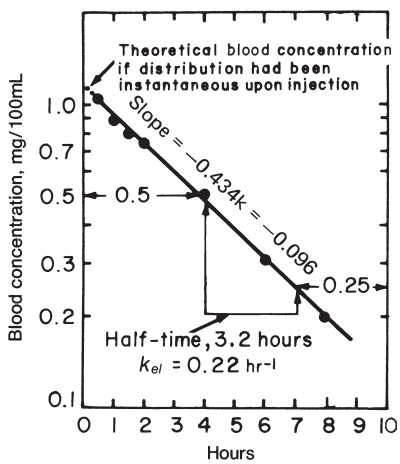


Figure 58-9. Semilog plot of the elimination curve in Figure 58-8. Note the log scale of the ordinate.

$$D_u = D_B \frac{k_u}{k_{el}}(1 - e^{-k_{el}t}) \quad (16)$$

k_u / k_{el} represents the fraction of the drug in the body that eventually is excreted in urine as unchanged drug and D_u^{∞} represents the total amount of unchanged drug excreted in urine. A plot of the log of the amount of drug remaining to be excreted ($D_u^{\infty} - D_u$) versus time yields a slope equal to $-k_e / 2.303$. This method requires collecting urine for at least 6 to 8 half-lives to achieve an accurate measure of D_u^{∞}.

For a drug eliminated by first-order kinetics, the elimination rate constant, k_{el}, can be expressed as the fraction of the volume of distribution that is presented to an eliminating organ and cleared of drug per unit time *(clearance)* relative to the total volume of distribution, V_D. Thus, k_{el} represents the fractional removal rate of drug from the system, and the elimination rate constant can be expressed in terms of clearance and volume of distribution:

$$\frac{\text{Clearance}}{V_D} = \text{Elimination-rate constant } (k_{el}) \quad (17)$$

As written in Equation 17, the elimination rate constant (and hence, plasma half-life) is a dependent parameter that, by itself, is not always the most reliable indicator of drug removal from the body. Disease or altered physiology (eg, aging, pregnancy) can alter protein binding, thereby affecting the apparent volume of distribution, or alter organ function, thereby affecting clearance, but these changes may not be reflected by changes in the half-life. For example, the volume of distribution may be altered due to changes in tissue or plasma protein binding, independent of specific organ function (clearance). In this instance, the half-life of a drug may be changed, but clearance could remain constant. Although a useful parameter, one must always bear in mind that half-life is a dependent or derived parameter that does not reliably reflect irreversible removal of drug from the body. A more accurate way to express half-life (Equation 14) therefore is

$$t_{1/2} = 0.693 \frac{V_D}{CL} \quad (18)$$

Clearance is the most useful pharmacokinetic indicator of irreversible loss of drug from the body and refers to a volume of fluid from which drug appears to be removed in a given amount of time. Clearance also can be expressed as the quotient of overall rate of elimination of a drug relative to the drug concentration at a particular organ of elimination,

$$\text{Clearance} = \frac{\text{Rate of elimination}}{\text{Concentration}} \quad (19)$$

and, if time-averaged over the time course of plasma concentrations, drug clearance can be expressed as

$$\text{Clearance} = \frac{\text{Amount of drug removed}}{AUC} \quad (20)$$

where AUC is the area under the concentration-versus-time curve. Total body clearance, CL_T, also can be estimated as the quotient of dose and area under the concentration-versus-time curve from zero to infinity.

$$CL_T = \frac{Dose_{IV}}{AUC_0^{\infty}} \quad (21)$$

Total body clearance is the sum of all the separate clearances that contribute to drug elimination

$$CL_T = CL_{RENAL} + CL_{HEPATIC} + CL_{OTHER} \quad (22)$$

INSTANTANEOUS INPUT WITH NONINSTANTANEOUS DISTRIBUTION

The one-compartment model adequately describes the pharmacokinetics of drugs with instantaneous distribution. However, for some compounds, distribution requires some finite time to reach equilibrium. During this time, the drug undergoes distribution and elimination, and drug concentrations decrease rapidly. When distribution equilibrium is established, the loss of drug from the body is due to elimination, and plasma concentrations decline more slowly. This biexponential time-course of plasma concentrations can be described by a two-compartment model. In this model, the body appears to behave as if it is comprised of two compartments, a central compartment and a peripheral compartment. By convention, drug absorption (or injection) and drug elimination occur from the central compartment and the peripheral compartment is closed and communicates with the environment only through the central compartment (Fig 58-10). The movement of drug between compartments following rapid intravenous injection with elimination from the central compartment can be described by

$$\frac{dD_1}{dt} = k_{21}D_2 - k_{12}D_1 - k_{10}D_1$$

and

$$\frac{dD_2}{dt} = k_{12}D_1 - k_{21}D_2 \quad (23)$$

where D_2 is the amount of drug in the peripheral or tissue compartment, D_1 is the amount of drug in the central compartment, k_{21} and k_{12} are the apparent first-order intercompartmental distribution rate constants, and k_{10} or k_{el} is the apparent first-order elimination rate constant from Compartment 1.

After intravenous injection of a drug that obeys two-compartment pharmacokinetics, the plasma concentration de-

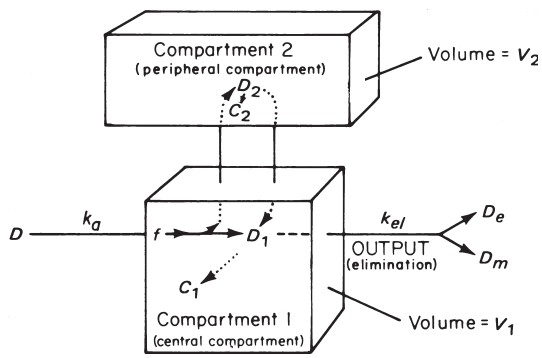


Figure 58-10. Diagram of open two-compartment pharmacokinetic model. The amount of dose that enters the body for an intravenous injection is the entire dose, and administration is instantaneous. The amount of dose absorbed from an extravascular dose is *F D,* where *F* is the fraction of dose absorbed with a rate constant, k_a. Some of the absorbed drug enters Compartment 2 with a first-order rate constant of k_{12} and is returned to Compartment 1 with a first-order rate constant of k_{21}. D_1 is the amount of drug in Compartment 1, and D_2 in Compartment 2; C_1 and C_2 are the respective concentrations in Compartments 1 and 2 ($C_1 = C_p$). Drug is eliminated from Compartment 1 with a first-order rate constant, k_{el}, which, however, is obscured by the lag in transfer of drug from Compartment 2 to Compartment 1. D_u is the amount excreted into urine, feces, expired air, sweat, milk, etc; D_m is the amount of drug metabolized. The relative volumes V_1 and V_2 may vary greatly, V_1 sometimes being the larger and other times the smaller.

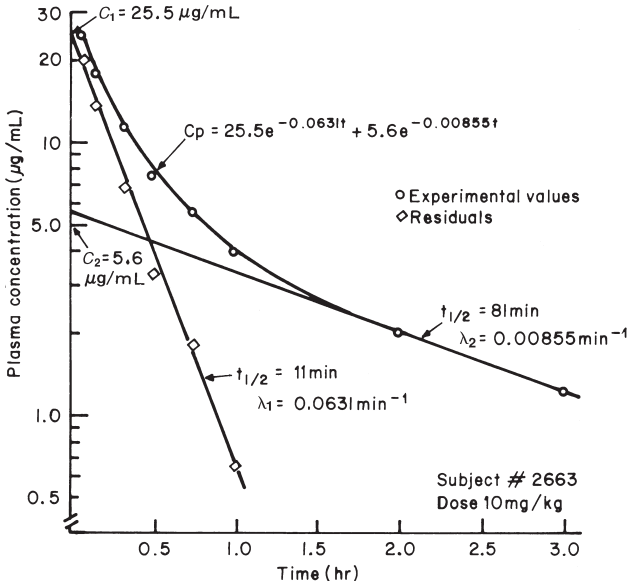


Figure 58-11. Resolution of the plasma concentration curve for pralidoxime into its distribution and elimination components after intravenous administration. Note that plasma concentration is plotted on a logarithmic scale. The time constant for the elimination phase is determined from the slope, $-0.434\ \lambda_2$; it is a hybrid constant and λ_2 is not the same as k_{el} (see text). Likewise, the time constant for distribution, λ_1, is obtained from the slope, $-0.434\lambda_1$, of the distribution line; λ_1 is also a hybrid constant. (From Gibaldi M, Perrier D. *Pharmacokinetics,* 2nd ed. New York: Dekker, 1982.)

clines in a complex biexponential fashion. When plotted on semilog graph paper, the separate processes of distribution and elimination can be identified by the method of residuals (Fig 58-11).[8] Figure 58-11 shows such a resolution of the biexponential decay into the two components of distribution and elimination. From the slopes and intercepts of the residuals, the plasma concentration, C, at any time, t, can be described as the sum of two exponentials, namely

$$C = C_1 e^{-\lambda_1 t} + C_2 e^{-\lambda_2 t} \tag{24}$$

where

$$C_1 = \frac{Dose(\lambda_1 - k_{21})}{V_1(\lambda_1 - \lambda_2)} \tag{25}$$

and

$$C_2 = \frac{Dose(k_{21} - \lambda_2)}{V_1(\lambda_1 - \lambda_2)} \tag{26}$$

Distribution is more rapid than elimination, such that at some point, the first term in Equation 24, $C_1 e^{\lambda_1 t}$ approaches zero and the biological half-life can be determined from the slope of the terminal phase

$$t_{1/2} = \frac{0.693}{\lambda_2} \tag{27}$$

$C1$, $C2$, λ_1, and λ_2 are hybrid constants, representing the intercepts, $C1$, $C2$, and slopes, λ_1, λ_2, of the two exponential phases, which can be obtained by computer-fitting the biexponential data. The zero-time intercept and the volume of the central compartment, V_1, and the actual pharmacokinetic parameters k_{12}, k_{21}, and k_{el} can be derived from the hybrid rate constants using the following relationships. At time $t = 0$

$$C_0 = C_1 + C_2;\ V_1 = \frac{Dose}{C_1 + C_2} \tag{28}$$

The hybrid rate constants, λ_1 and λ_2, can be defined using the following two equations:

$$\lambda_1\lambda_1 = k_{21}k_{el} \tag{29}$$

$$\lambda_1 + \lambda_1 = k_{12} + k_{21} + k_{el} \tag{30}$$

Thus

$$k_{el} = \frac{\lambda_1\lambda_2}{k_{21}} \tag{31}$$

$$k_{21} = \frac{C_1\lambda_2 + C_2\lambda_1}{\lambda_1 - \lambda_2} \tag{32}$$

$$k_{12} = \lambda_1 + \lambda_2 - k_{21} - k_{el} \tag{33}$$

The reader is referred to Gibaldi and Perrier (see *Bibliography*) for a more in-depth derivation of these expressions. The slope of the final phase of biexponential disposition, λ_2, can be related to the elimination-rate constant, k_{el}, by

$$\lambda_2 = f_C k_{el} \tag{34}$$

where f_C is the fraction of the drug that is in the central compartment after distribution equilibrium has been achieved. After distribution, the fraction of the drug in the central compartment is a constant.

$$f_C = \frac{k_{21} - \lambda_2}{k_{21} + k_{12} - \lambda_2} \tag{35}$$

The terminal disposition constant, λ_2, reflects disposition from the entire body and is a function of distribution and elimination. The rate constant, k_{el}, represents only elimination from the theoretical central compartment.

The volume of distribution, V_D, can be determined in a two-compartment system; however, the estimation is complicated by the noninstantaneous nature of the distribution phase between the two compartments and results in the apparent volume of distribution being time-dependent. From Equation 18, it can be seen that the volume of distribution of the central compartment, V_1, can be obtained following administration of an intravenous dose, D_{IV}, of drug from

$$V_1 = \frac{D_{IV}}{C_1 + C_2} \tag{36}$$

Clearance can be calculated from the product of k_{el} and V_1, and the volume of the central compartment can be expressed as

$$V_1 = \frac{D_{IV}}{k_{10}[AUC]_{0\to\infty}} \tag{37}$$

The most accurate method of estimating the volume of distribution is to estimate the steady-state volume of distribution. The volume of distribution at the steady state, V_{SS}, represents the steady state with respect to distribution of the drug from the central compartment to the tissue compartments and is not altered by changes in drug elimination or clearance. The total amount of the drug in the body at the steady state is the sum of the amounts in all compartments, thus

$$V_{SS} = V_1 + \frac{k_{12}}{k_{21}} V_1 \tag{38}$$

Notice that V_{SS} is independent of the elimination rate constant, k_{el}, and λ_2.

The volume of distribution by area, V_β, is an alternate method of estimating the apparent volume of distribution and relies on

the assumption that the plasma and the amount of drug in the body decline in parallel during the postdistributive phase.

$$V_\beta = \frac{V_1 k_{el}}{\lambda_2} \tag{39}$$

The least accurate method of estimating the volume of distribution for a drug that follows biexponential elimination kinetics is by extrapolation, V_{EXTRAP}, because changes in distribution can alter the estimation of the hybrid intercept, C_2.

$$V_{EXTRAP} = \frac{D_{IV}}{C_2} \tag{40}$$

Distribution to various tissues depends upon both blood flow to that tissue and the rate of uptake (effective partition coefficient) into a particular tissue and its cells. The overall pattern of drug distribution is governed by both tissue perfusion and diffusion of drug within tissues. Tissues with the highest blood flow, such as liver, kidney, and brain, equilibrate more rapidly than tissues that are perfused less well, such as skin and fat. Once in the tissue vasculature, drug distribution into tissue is controlled largely by diffusional barriers of cell membranes. Rowland and Tozer (see *Bibliography*) present a useful expression for the first-order rate constant for distribution of drug into tissue.

$$k_{TISSUE} = \frac{(Q/V_{TISSUE})}{k_{PARTITION}} \tag{41}$$

where $k_{PARTITION}$ is the equilibrium distribution ratio of tissue and venous drug concentration, Q is tissue blood flow, V_{TISSUE} is the tissue volume, and the quotient of Q and V_{TISSUE} is the tissue perfusion rate. The time to reach tissue equilibrium is the reciprocal of Equation 41. For a poorly perfused tissue such as fat, the $k_{PARTITION}$ may be quite high and the Q/V_{TISSUE} low, resulting in a long time to reach tissue equilibrium. Even for highly perfused tissues, such as the brain, the distribution of some drugs may be quite variable and will depend upon diffusion across cell membranes. In this case, distribution is said to be diffusion-rate-limited and will depend upon both the oil-to-water partition coefficient and the degree of ionization at physiological pH.

For most drugs, distribution usually occurs more rapidly than elimination, resulting in complete distribution before most of the drug has been eliminated. For some drugs, once injected, distribution is so rapid that the overall plasma-concentration time-course represents elimination (see Fig 58-7). Thus, both administration and distribution appear to be instantaneous, and the pharmacokinetics can be modeled by the simplest one-compartment model (see Fig 58-7). For such a drug, the volume of distribution can be calculated as the quotient of the intravenous dose and the extrapolated plasma concentration at time zero, C_0.

CONTINUOUS INPUT

It is sometimes desirable to administer a drug continuously to maintain constant plasma concentration. This is often the case for drugs with very rapid elimination or for those that have a low therapeutic index. Continuous input commonly is thought of in terms of intravenous infusion; however, sustained-release oral dosage forms and delivery of drugs through the skin from patches also are examples of continuous input, and the pharmacokinetics of drug administration is similar for all systems with continuous input.

With constant intravenous infusion, the plasma concentration rises in a logarithmic fashion and eventually reaches a plateau (Fig 58-12).[8] The time to reach the plateau or steady-state concentration is determined by the elimination rate constant. The rate of change of drug in the body (D_B) during a constant rate infusion (R_0) is the difference between the zero-order infusion rate and the first-order elimination rate.

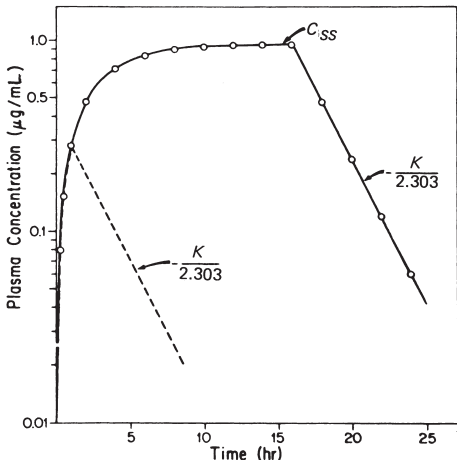


Figure 58-12. Semilogarithmic plot of plasma concentration during and after cessation of a constant intravenous infusion of a drug in a one-compartment system. Whether infusion is stopped prior to the attainment of a plateau or after, the plasma concentration will fall log-linearly with a slope of $-0.434k_{el}$. In the figure, K is k_{el} and 1/2.303 = 0.434. C_{ss} is the steady-state concentration, C_p^{ss}. (From Gibaldi M, Perrier D. *Pharmacokinetics,* 2nd ed. New York: Dekker, 1982.)

$$\frac{dD_B}{dt} = R_0 - (k_{el} \cdot D_B) \tag{42}$$

The plasma concentration (C) at any time during the constant infusion is

$$C = \frac{R_0}{CL_T}(1 - e^{-k_{el}t_{inf}}) \tag{43}$$

where t_{inf} equals the time of the infusion. As the time of the infusion increases, the exponential expression approaches zero, and the concentration approaches steady state. At steady state, the rate of infusion is equal to the rate of elimination, and the simplified expression can be expressed as

$$C_{SS} = \frac{R_0}{CL_T} \tag{44}$$

The fraction of the steady state that is achieved in some time, t_{inf}, after the start of the infusion can be calculated as

$$\frac{C}{C_{SS}} = (1 - e^{-k_{el}t_{inf}}) \tag{45}$$

and can be expressed in terms of half-lives as

$$\frac{C}{C_{SS}} = (1 - 2^{-n}) \tag{46}$$

where n is the ratio of infusion time and half-life. For example, when the infusion time equals three half-lives ($n = 3$), the concentration is at 87.5% of the steady state, and when the infusion has lasted for four half-lives ($n = 4$), the concentration has achieved 93.75% of the steady state. Theoretically, one never reaches steady-state conditions because this is an exponential process; however, for clinical purposes one can assume, with little error, that steady-state concentrations are achieved within four to five half-lives.

If a drug has a relatively long half-life and the therapeutic situation demands rapid attainment of therapeutic plasma concentrations, it is sometimes desirable to give a loading dose at the beginning of the constant-rate infusion. The loading dose should approximate the amount of drug in the body at steady state. If the apparent volume of distribution and target concentration is known, the loading dose can be calculated simply as

$$Dose_{LOADING} = (\text{Target concentration})(V_D) \tag{47}$$

The plasma concentration is the sum of the contributions from the loading dose and the infusion and can be estimated at any time after the loading dose has been given and infusion started from

$$C = \frac{Dose_{LOADING}}{V_D}(e^{-k_{el}t}) + \frac{R_0}{CL_T}(1 - e^{-k_{el}t_{inf}}) \quad (48)$$

and if the half-life of the drug is known, the loading dose can be estimated from the quotient of infusion rate, R_0, and elimination rate constant, k_{el}. For some drugs, such as lidocaine, the entire loading dose cannot be given in a single bolus injection because there is a significant distribution phase. In such a case, fractional loading-dose schemes can be used in which the loading dose is divided into several smaller bolus doses and given during the beginning of the infusion.

Finally, it should be noted that whether or not a loading dose is given, the attainment of steady state is determined by the elimination half-life and not by the rate of the infusion or the use of bolus loading doses to achieve concentrations rapidly.

MULTIPLE-DOSE ADMINISTRATION

Continuous administration of a drug is often impractical, and multiple-dose regimens are used to maintain the concentration of a drug within an acceptable range that mimimizes the development of toxicity and avoids loss of efficacy. Usually, the dose of a drug is administered with a constant dose interval, referred to as τ. Some features of a multiple dosage scheme are shown in Figure 58-13. The drug is administered at a fixed dose and a fixed interval. Each successive dose is administered before the previous dose has been eliminated entirely, and thus drug accumulation occurs. As with the constant intravenous infusion, the time to reach a steady-state fluctuation depends upon the elimination half-life and not on the size of the dose or the dosing interval. In Figure 58-13, the dose, D, is administered at a dosing interval equal to the half-life. After the first dose is given, the amount of drug in the body is equal to that dose. When the next dose is given, the amount of drug in the body is equal to $D + 0.5D$. At the end of each dose interval, the total amount of drug in the body is half of the postinjection peak and is the sum of the amount remaining from all of the previous doses. The maximum, $C_{MAX,SS}$, and minimum, $C_{MIN,SS}$, concentrations at steady state are described by

$$C_{MAX,SS} = \frac{Dose/V_D}{(1 - e^{-k_{el}\tau})} \quad (49)$$

and

$$C_{MIN,SS} = C_{MAX,SS}e^{-k_{el}\tau} \quad (50)$$

The concentration at the midpoint of a dosing interval at the steady state is a time-averaged concentration, C_{AV}, over the entire dosing interval and is described by

$$C_{AV} = \frac{D_{IV}}{V_D k_{el}\tau} = \frac{1.44 t_{1/2} D_{IV}}{\tau} \quad (51)$$

NONCOMPARTMENTAL ANALYSIS FOLLOWING INSTANTANEOUS INPUT

There are numerous disadvantages associated with viewing drug disposition from a compartmental perspective, not the least of which is the lack of physiological relevance of such models. Noncompartmental models have been developed and are generally the *preferred* method for assessing overall drug disposition, in part because parameters such as volume of distribution and clearance can be calculated directly from the data, without computer fitting. Moreover, parameters estimated by these methods generally are less sensitive to variability in the data. Noncompartmental methods characterize drug disposition using time- and concentration-averaged parameters and have been described by Jusko.[9] This analysis also is known as nonparametric analysis and assumes that all processes are first-order and that the parameters of the model reflect steady-state behavior. A primary tool used in this methodology is that of statistical moments.[10]

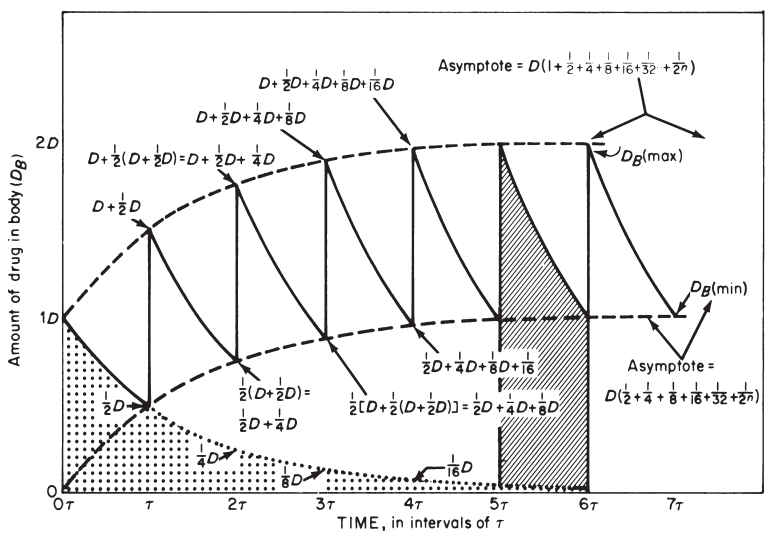


Figure 58-13. The accumulation of drug in the body during a regimen of multiple dosing. Dose, D, is administered intravenously at intervals, τ, equal to the half-life, $t_{1/2}$. Thus, after each dose, the amount in the body, D_B, has decreased to half the previous peak amount at the time each dose is administered. When the cumulated amount in the body after injection reaches $2D$, the body content will fluctuate from $2D$ to $1D$ during each dose interval thereafter. Approximately five half-lives are required before this leveling off (plateau) of the body content occurs. The stippled area is the area under the elimination curve of a single injection, if no second dose had been given. The cross-hatched area is the area under the curve during a single-dose interval at steady state. The two areas are equal.

STATISTICAL MOMENTS—The use of statistical moments in the analysis of the time-course of drug concentrations is especially useful because it frees the investigator from the use of such models as the compartmental, which often are derived empirically and do not represent physiological events.

The time-course of drug concentration in blood generally can be viewed as a statistical distribution curve and described in a similar manner as any other array of data. A moment is simply a mathematical description of a discrete distribution of data. In the field of statistics, for example, the sample size (n), mean, and variance are the zero (M_0), first (M_1) and second (M_2) moments, respectively, for an array of data. In physics, for example, weight, center of mass, and moment of inertia represent (M_0), (M_1), and (M_2), respectively.

In statistics, the mean of a population is estimated by the sample mean. Similarly, in pharmacokinetics one may calculate estimates of the true function that describes the drug concentration versus time using statistical moments. Assume the existence of a theoretical relationship for $C(t)$ as a function of time. The nonnormalized moments, S_r, where $r = 0, 1, 2, \ldots$, m^{th} moment, about the origin are calculated as

$$S_r = \int_0^\infty t^r C(t)\, dt \quad (r = 0, 1, 2, \ldots m) \tag{52}$$

Hence

$$S_0 = \int_0^\infty C(t)\, dt = AUC \tag{53}$$

$$S_1 = \int_0^\infty tC(t)\, dt = AUMC \tag{54}$$

where $AUMC$ is the area under the $C \cdot t$ versus time curve, whereas S_0 and S_1 are the zero and first nonnormalized moments, respectively. These two parameters, AUC and $AUMC$, are derived from the drug concentration versus time data and are used to calculate the pharmacokinetic parameters of interest.

The use of noncompartmental methods requires a means of determining the AUC. While several methods are available for such determinations, the simplest is the use of the trapezoidal rule.[11] This permits calculation of the AUC or the $AUMC$ from zero to the time of the last sample (t^n). However, it generally is necessary to determine the area from zero to infinity. If one assumes that there is log-linear decline from t^n to infinity, then

$$AUC_{t^n}^{\infty} = \int_{t^n}^{\infty} C\, dt = \frac{C^n}{\lambda_Z} \tag{55}$$

where λ_Z is the slope of the terminal exponential. Thus, the AUC from zero to infinity can be calculated as

$$AUC_0^{\infty} = AUC_0^{t^n} + \frac{C^n}{\lambda_Z} \tag{56}$$

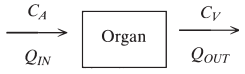
for the $AUMC$

$$AUMC_{t^n}^{\infty} = \frac{C^n}{(\lambda_Z)^2} + \frac{t^n C^n}{\lambda_Z} \tag{57}$$

and

$$AUMC_0^{\infty} = AUMC_{\infty}^{t^n} + \frac{C^n}{(\lambda_Z)^2} + \frac{t^n C^n}{\lambda_Z} \tag{58}$$

PHARMACOKINETIC PARAMETERS DERIVED FROM STATISTICAL MOMENTS—There are four parameters of primary interest that are derived using statistical moments, the most important of which is clearance. A conceptual consideration of clearance is provided by considering an organ through which blood containing a drug flows.



where Q_{IN}, Q_{OUT}, C_A, and C_V are the blood flow in and out and the concentration of the drug in arterial and venous blood, respectively. If $C_V < C_A$, the organ is capable of elimination and is referred to as a clearing organ, or an organ of elimination. The elimination of a drug can be described using mass balance considerations:

$$\text{Rate of the drug entering the organ} = QC_A \tag{59}$$

$$\text{Rate of the drug leaving the organ} = QC_V \tag{60}$$

$$\text{Rate of elimination of the drug} = QC_A - QC_V = Q(C_A - C_V) \tag{61}$$

The ratio of the rate of drug elimination and the rate at which the drug enters the organ is defined as the *extraction ratio, E*.

$$E = \frac{\text{Rate of elimination}}{\text{Rate of entry}} = \frac{Q(C_A - C_V)}{QC_A} = \frac{(C_A - C_V)}{C_A} \tag{62}$$

The extraction ratio is a measure of the efficiency with which an organ eliminates a drug. From this parameter, the organ clearance of a drug can be described as

$$CL_T = QE = \frac{Q(C_A - C_V)}{C_A} \tag{63}$$

Recall that clearance is defined as the volume of blood from which all of the drug would appear to be removed per unit time. By analogy to the definition of organ clearance, one can define the total or systemic clearance as the ratio of overall elimination rate, dX/dt, to drug concentration in blood, C:

$$CL_T = \frac{dX/dt}{C} \tag{64}$$

Integrating from zero to infinity yields

$$CL_T = \frac{\int_0^\infty \frac{dX}{dt}\, dt}{\int_0^\infty C\, dt} \tag{65}$$

where the numerator is the total amount of the drug ultimately eliminated (the IV dose) and the denominator is the AUC from zero to infinity. Thus, the total clearance is the quotient of intravenous dose and the AUC from zero to infinity.

$$CL_T = \frac{Dose_{IV}}{AUC_0^{\infty}} \tag{66}$$

The volume of distribution at the steady-state, V_{SS}, most reliably measured during a steady-state infusion, now can be determined using data from single-dose experiments and employing statistical moment analysis.[12]

$$V_{SS} = \frac{(Dose_{IV})(AUMC)}{(AUC)^2} \tag{67}$$

Equation 67 assumes that

1. All processes involved in drug disposition (eg, distribution, elimination) are linear.
2. The drug is administered to and eliminated via the sampling site.
3. There is instantaneous input.

If the drug is administered via a short infusion, the volume of distribution at the steady state can be estimated from

$$V_{SS} = \frac{(R_0 T)(AUMC)}{(AUC)^2} - \frac{R_0 T^2}{2(AUC)} \tag{68}$$

where R_0 is the rate of infusion and T is the duration of the infusion.

Another important pharmacokinetic parameter that can be determined using statistical moment analysis is the *systemic availability, F,* which is a measure of the fraction of the administered dose that reaches the systemic circulation following oral administration. This parameter can be calculated as

$$F = \frac{AUC_{PO}Dose_{IV}}{AUC_{IV}Dose_{PO}} \tag{69}$$

where AUC_{PO} and $Dose_{PO}$ are the oral area under the concentration-versus-time curve and oral dose, respectively.

When administering a drug, the amount administered in terms of gross weight (eg, mg, g, or μ g) often is considered. It is, however, probably more appropriate to focus on *molecules* when considering pharmacokinetic and pharmacodynamic events. Even the administration of a relatively small dose of drug may represent a large number of molecules. Consider the administration of 1 mg of a drug with a molecular weight of 300 daltons. The number of molecules in this dose is approximately 2×10^{18}. Instantaneous administration of the entire dose will result in drug molecules spending various amounts of time in the body. After the intravenous injection of a drug, one can imagine that some of the drug molecules are eliminated immediately, whereas some of the molecules require a longer time to be eliminated, and some molecules even require a very long time to be eliminated. The time spent in the body, for a given molecule, is its residence time. The *mean residence time, MRT,* is the sum of all the residence times divided by the number of molecules. A conceptual understanding of this can be gained from the following example.

Assume a child receives 20 dimes for his birthday and immediately places them in his piggy bank. Over the next month, he periodically removes one or more dimes from the piggy bank to purchase candy. Specifically, 3 days after placing the coins in his bank, he removes 5 dimes, on day 10 he removes 4 dimes, on day 21 he removes 6 dimes, and on day 30 he removes 5 dimes. At the 30th day after placing the coins in his bank, all of the coins have been removed. Hence, the *elimination* of dimes from the bank is complete. The *MRT* of dimes in the piggy bank is simply the sum of the times that coins spend in the bank divided by the number of dimes placed in the bank:

$$MRT = \frac{3 + 3 + 3 + 3 + 3 + 10 + 10 + 10 + 10 + 21 + 21 + 21 + 21 + 21 + 21 + 30 + 30 + 30 + 30 + 30}{20}$$

$$MRT = \frac{3*5 + 10*4 + 21*6 + 30*5}{20}$$

$$MRT = 16.55 \text{ days}$$

This provides a relationship with which one can determine the *MRT* for any given number of drug molecules, A_i, which spend a given amount of time in the body of, t_i, thus

$$MRT = \frac{\sum_{i=1}^{n} A_i t_i}{A_{TOTAL}} \tag{70}$$

where n equals the total number of residence times. The mean rate of drug leaving the body relative to the total amount eliminated also can be expressed in terms of concentration.

$$MRT = \frac{\int_0^\infty tC(t)\,dt}{\int_0^\infty C(t)\,dt} = \frac{AUMC}{AUC} \tag{71}$$

Equation 71 is not a definition of *MRT,* rather it is the derived expression from which one can calculate *MRT* when clearance is constant. The mean residence time assumes instantaneous administration, and therefore, it is technically incorrect to calculate the mean residence time following an oral dose using the quotient of $AUMC_{PO}$ and AUC_{PO}. When calculated in this manner, it often is stated that the *MRT* is a function of the route of administration. Actually, *MRT* is *independent* of the route of administration because the mean time that molecules reside in the body is not influenced by the route of administration.[13] However, the interpretation of the ratio of *AUMC* and *AUC* does change as a function of administration because this ratio only yields the *MRT* when the input is instantaneous.

A better way to express the route dependence of the *AUMC*/*AUC* is to refer to this ratio as the *mean transit time, MTT*. The *MTT* is the average time required for drug molecules to leave a kinetic system after administration. Thus, because an IV bolus assumes instant input,

$$AUMC_{IV}/AUC_{IV} = MRT = MTT_{IV} \tag{72}$$

whereas

$$AUMC_{PO}/AUC_{PO} = MTT_{PO} = MTT_{IV} + MAT = MRT + MAT \tag{73}$$

where *MAT* is the mean absorption time. Thus, for oral absorption, the *AUMC*/*AUC* provides an *MTT* in the kinetic system that is composed of the gastrointestinal (GI) tract and the body. The *MRT* also can be calculated as the quotient of the V_{SS} and clearance. Finally, one can relate the *MRT* to the elimination half-life by considering the situation in which a drug displays monoexponential decline. The *MRT* can be written as

$$MRT = \frac{AUMC}{AUC} = \frac{C_0/\lambda^2}{C_0/\lambda} = \frac{1}{\lambda} \tag{74}$$

and represents the time required for 63.2% of an intravenous dose to be eliminated from the body.

ABSORPTION

If a drug is administered intravenously in a single, rapid injection, the process of absorption is bypassed. The time for this injection is typically so short compared with other pharmacokinetic processes that it is ignored. As previously described for a one-compartment model, peak plasma concentration and distribution equilibrium are achieved instantaneously. This is depicted in Figure 58-14A.[14] In the model for the figure, there is no elimination, and the concentration remains constant following administration. With a constant intravenous infusion (*B*), the concentration rises rectilinearly so long as the infusion is maintained at a constant, zero-order rate. With other routes of administration, there are delays in the appearance of drug in the vascular system because the drug must be absorbed from the site of administration (oral, intramuscular, subcutaneous, rectal). Drug absorption depends upon both the physicochemical properties of the drug (pK_a, dosage form, partition coeffi-

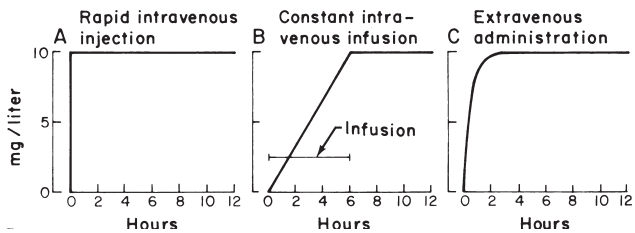


Figure 58-14. Time-concentration curves for injection *(A),* infusion *(B),* and extravenous *(C)* administration of drug in the one- compartment model. The volume of the compartment is 100 L (V_d = 100 L); the amount of drug administered in each instance is 1000 mg. Drug elimination has been set to zero, so that the time- concentration curve for each model of administration can be examined without the complication of simultaneous elimination. (Adapted from Bigger JT. *Am J Med* 1975; 58:479.)

cient) and the physiology of the site of absorption (surface area, blood flow). Most drugs are absorbed by simple diffusion, and the kinetics are first-order. Zero-order absorption occurs for some processes that are saturable and for sustained-release dosage forms. Absorption and elimination of a drug are a sequential process, and the rate of change of drug in the body is the difference between the rate of uptake (absorption) and rate of efflux (elimination). For a drug that is absorbed by a first-order process and eliminated by a first-order process, with instantaneous distribution, the rate of change of the amount of drug in the body can be expressed as

$$\frac{dD_B}{dt} = RATE_{IN} - RATE_{OUT} \tag{75}$$

For a drug that is absorbed from the GI tract, the rate of change of the amount of drug in the body is

$$\frac{dD_B}{dt} = Fk_aD_{GI} - kD_{BODY} \tag{76}$$

The time-course of absorption and elimination is shown in Figure 58-15.[7] The plasma concentration at any time t is equal to

$$C = \frac{k_aD_0F}{V_D(k_a - k_{el})}(e^{-k_{el}t} - e^{-k_at}) \tag{77}$$

where F is the fraction of the dose, D_0, that is absorbed from the GI tract and k_{el} and k_a are the first-order rate constants for elimination and absorption, respectively. The time to reach the maximum concentration, t_{MAX}, can be determined from

$$t_{MAX} = \frac{2.3 \log (k_a/k)}{k_a - k} \tag{78}$$

and this time substituted into Equation 77 will determine the maximum concentration, C_{MAX}. The rising phase of the plot (see Fig 58-15) is not log-linear because absorption and elimination are occurring simultaneously. At C_{MAX}, the absorption rate is equal to the elimination rate, and after absorption is complete, the plot declines in a log-linear manner. This log-linear line described by the elimination phase, when extrapolated to zero time, yields a theoretical zero-time concentration. The absorption rate constant, k_a, can be obtained from the difference between the empirical curve and the extrapolated line using the

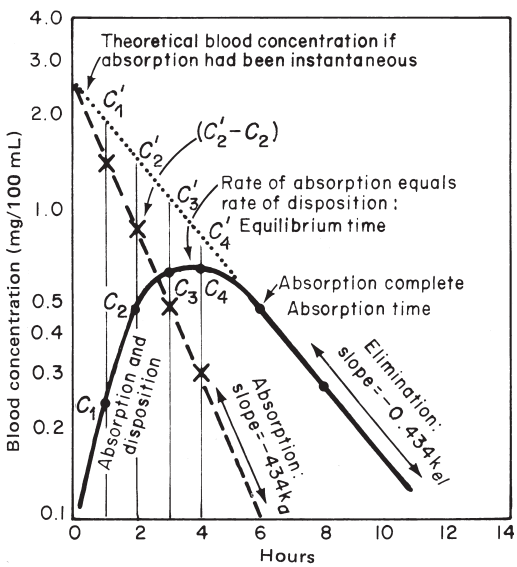


Figure 58-15. Kinetics of absorption and disposition of theophylline in a human subject after oral administration of 0.5 g of aminophylline per 70 kg. Blood concentration is plotted on a log scale. (Data from Truitt EB Jr, et al. *J Pharmacol Exp Ther* 1950; 100:309.)

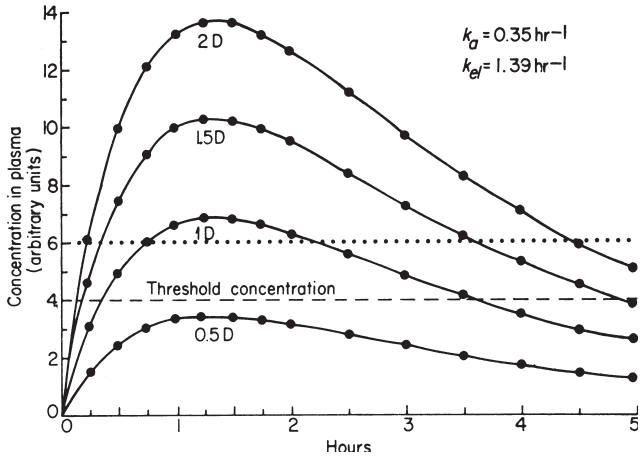


Figure 58-16. The effect of the size of the dose of a drug on the peak concentration, time of peak concentration, and duration of action. The data were calculated from a one-compartment model.

method of residuals. This is a commonly used technique in pharmacokinetics to separate a curve into its component parts and is often referred to as *feathering, stripping,* or *peeling* the curve. The reader is referred to Gibaldi and Perrier (see *Bibliography*) for a more comprehensive discussion with examples of the application of this technique.

That the peak concentration should vary with the dose is self-evident from Equations 77 and 78 and from Figure 58-16. The time of the peak concentration is the same for all doses. The time to peak concentration can be affected by both the absorption rate and the elimination rate. In Figure 58-17, the effect of altering the absorption rate on the time to peak concentration is shown. With faster absorption, the time to peak concentration occurs earlier and is higher than with slower absorption. Figure 58-18 shows the effect of altering the elimination rate constant on the t_{MAX}. With a rapid elimination rate (shorter half-life), the peak concentration occurs sooner and is lower than with slower elimination (longer half-life).

The maximum concentration at the steady state for an oral regimen is given by

$$C_{MAX,SS} = \frac{FDose}{V_D}\left(\frac{1}{1 - e^{-k\tau}}\right)e^{-kt_{PEAK}} \tag{79}$$

The minimum concentration at the steady state is

$$C_{MIN,SS} = \frac{k_aFD}{V(k_a - k)}\left(\frac{1}{1 - e^{-k\tau}}\right)e^{-k\tau} \tag{80}$$

To design multiple oral dosing regimens, the equations that describe the maximum and minimum concentrations at the steady state are somewhat unwieldy. In clinical practice, values for k_a are not always readily available, and in such instances the equations for the maximum and minimum concentrations at the steady state following intravenous administration will generally suffice as long as one recalls that because absorption is not instantaneous, the peak concentration and time to peak concentration will not occur immediately.

ORGAN-SPECIFIC CLEARANCE

The total clearance of a drug from the body almost always involves more than one organ of elimination. The anatomy of the human body is such that the clearance from the composite clearing organs occurs in parallel and is, therefore, additive. For example, if a drug is eliminated solely by hepatic and renal elimination, the total clearance, CL_T, of the drug is given as

$$CL_T = CL_H + CL_R \tag{81}$$

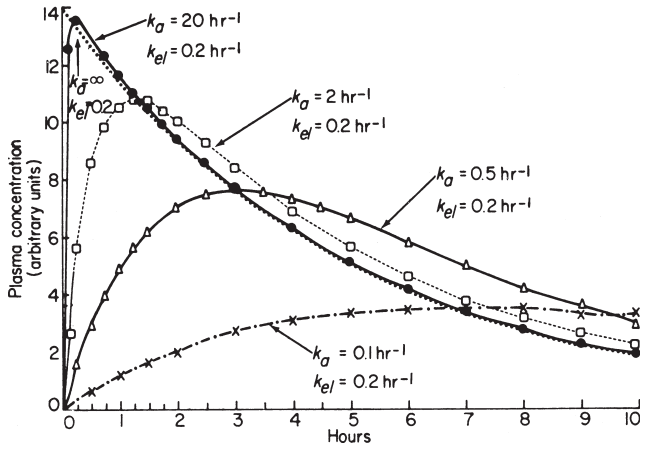


Figure 58-17. The effect of differences in the rate of absorption of drugs on the peak concentration, time of peak concentration, and sojourn in the body. The rate of elimination is the same for all curves. The dotted line ($k_a = \infty$) is approximately what the concentration curve would be, had the drug been given intravenously. The data were calculated from a one-compartment model.

where CL_H and CL_R are the hepatic and renal clearance, respectively. Measurement of the total amount of drug excreted unchanged in urine, D_U, after an intravenous dose, D_{IV}, allows the calculation of the fraction of the drug eliminated renally, F_r, where

$$F_r = \frac{D_U}{D_{IV}} \tag{82}$$

The renal clearance, CL_R, may be determined as the product of total clearance and the fraction of the drug eliminated by the kidney. If the liver is the only other eliminating organ, the hepatic clearance is given by

$$CL_H = CL_T - CL_R \rightarrow (1 - F_r)CL_T \tag{83}$$

One exception to the principle of the additivity of organ clearances is pulmonary drug elimination. This exception is because the lung is in circulatory series with the rest of the body organs such that 100% of cardiac output passes through the lungs. Few drugs exhibit significant elimination by the lungs so that this exception is rarely of concern in the overall assessment of drug elimination.

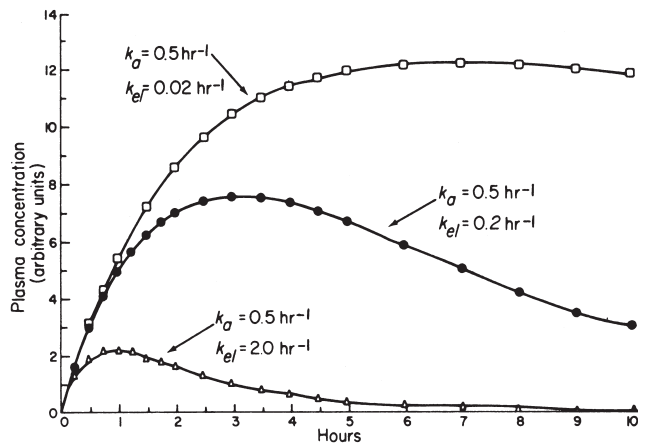


Figure 58-18. The effect of differences in the rate of elimination of drugs on the peak concentration, time of peak concentration, and sojourn in the body. The rate of absorption is the same for all curves. The data were calculated from a one-compartment model.

HEPATIC CLEARANCE—It was shown previously that the ratio of the rate of drug elimination and the rate at which drug enters the organ of elimination is defined as the extraction ratio, E, and is a measure of the efficiency with which an organ eliminates a given drug. One can define the organ clearance of a drug as the product of blood flow to the organ, Q, and the extraction ratio, and for hepatic clearance the equation becomes

$$CL_H = Q_H E \tag{84}$$

where Q_H is the hepatic blood flow. While an initial examination of this simplistic model for hepatic clearance would suggest that CL_H is directly proportional to Q_H, this conclusion is not correct because E varies inversely with Q_H. Specifically, as Q_H increases, E decreases. This observation indicates that a more complex model of hepatic clearance is necessary if quantitative and qualitative predictions of hepatic drug clearance are to be made. In particular, this parameter must be described in terms that are physiologically independent.

Numerous models have been proposed and tested to describe the hepatic clearance of drugs. While a discussion of the advantages and disadvantages of the various models proposed is beyond the scope of this chapter, the *venous equilibrium* model of hepatic clearance has shown substantial utility in the prediction of both pathophysiological and drug-induced changes in hepatic clearance. For a good discussion of the various models of hepatic clearance, see the review by Morgan and Smallwood.[15] In the venous equilibrium model, the hepatic extraction is described by

$$E = \frac{f_{ub}CL_{u,int}}{Q_H + f_{ub}CL_{u,int}} \tag{85}$$

where f_{ub} and $CL_{u,int}$ are the unbound fraction in blood and the unbound intrinsic hepatic clearance, respectively. The unbound intrinsic clearance reflects the ability of the liver to remove drug from blood in the absence of other confounding factors, such as Q_H and f_{ub}. Since it has already been shown that hepatic clearance is the product of Q_H and E

$$CL_H = \frac{(Q_H)(f_{ub}CL_{u,int})}{Q_H + f_{ub}CL_{u,int}} \tag{86}$$

This model for hepatic clearance provides a powerful tool for predicting changes in drug clearance and, subsequently, steady-state drug concentrations, when certain limiting conditions are met. In particular

When $Q_H >> f_{ub}\ CL_{u,int}$, CL_H

can be approximated by $f_{ub}\ CL_{u,int}$ (87)

When $Q_H << f_{ub}\ CL_{u,int}$, CL_H can be approximated by Q_H (88)

Compounds with a high $f_{ub}CL_{u,int}$ are said to exhibit perfusion rate–limited elimination; that is, their elimination rate will be rate-limited by hepatic blood flow. Compounds with a low $f_{ub}CL_{u,int}$ are said to be perfusion rate-independent. These limiting conditions allow us to place many drugs into classifications that exhibit similar pharmacokinetics. For example, agents with an $f_{ub}CL_{u,int} < 0.2$ L/min can be classified as low intrinsic clearance drugs, whereas those with an $f_{ub}CL_{u,int} > 5$ L/min are defined as exhibiting a high intrinsic clearance (Table 58-1).

The venous equilibrium model also serves as a useful tool in the assessment of the impact of changes in protein binding on hepatic clearance. Recall in Equation 87 that for a drug exhibiting a low intrinsic clearance, changes in protein binding will result in proportional changes in hepatic clearance. This type of drug is said to exhibit *restrictive clearance;* that is, only the free (or unbound) drug is available for clearance by the liver. High intrinsic clearance drugs, on the other hand, are said to exhibit *nonrestrictive clearance*.

These relationships provide important insight into the effect of changes in protein-binding on the steady-state concentration

Table 58-1. Examples of Drugs with Low and high Intrinsic Clearances That Are Eliminated Largely by Hepatic Metabolism

LOW $f_{ub}CL_{u,int}$ (<0.2 L/min)	HIGH $f_{ub}CL_{u,int}$ (>5.0 L/min)
Antipyrine	Chlorpromazine
Barbiturates	Encainide
Diazepam	Meperidine
Digitoxin	Metoprolol
Isoniazid	Organonitrates
Phenytoin	Propafenone
Theophylline	Propranolol
Tolbutamide	Tricyclic antidepressants
Warfarin	Verapamil

of drugs. Consider the case of a drug being administered as a constant-rate intravenous infusion. As described previously

$$C_{ss} = R_0/CL_T \rightarrow R_0/CL_H \tag{89}$$

for a drug solely eliminated by the liver. In the case of a drug with a low intrinsic clearance, Equation 89 can be simplified to

$$C_{ss} = R_0/(f_{ub}CL_{u,int}) \tag{90}$$

If f_{ub} were to be increased, for example, by displacement from protein-binding sites, the steady-state concentration would decrease. This may lead one to the conclusion that the infusion rate needs to be increased to maintain the original steady-state concentration. However, one needs to examine the effects of altered physiology on the free or unbound drug concentration, $C_{u,ss}$

$$C_{u,ss} = f_{ub}C_{ss} \rightarrow C_{ss} = C_{u,ss}/f_{ub} \tag{91}$$

substituting for C_{ss} in Equation 90 and solving for $C_{u,ss}$ yields

$$C_{u,ss} = R_0/CL_{u,int} \tag{92}$$

It can be seen that the steady-state concentration of unbound (active) drug is independent of changes in the free fraction, and no dosage adjustment would be necessary. This conclusion also is valid following the oral administration of a drug with a low intrinsic clearance.

In contrast, for a drug with a high intrinsic clearance, a change in f_{ub} will result in a proportional change in the steady-state unbound drug concentration during a constant-rate infusion.

An additional consideration, which must be accounted for with high intrinsic clearance drugs, is the impact of the first-pass effect. When a drug is absorbed from the stomach and small intestine, the venous blood from the sites of absorption enters into the portal venous flow. This results in all of the absorbed drug passing through the liver prior to entry into the systemic circulation. For drugs that exhibit a high intrinsic clearance, the consequence of presystemic hepatic metabolism is a substantial reduction in the systemic availability of the drug when administered orally. This phenomenon explains the marked discrepancy between an oral and an intravenous dose of a given drug required to achieve identical plasma concentrations. For example, the therapeutic dose of propranolol ranges between 1 and 6 mg intravenously, whereas the oral doses necessary to achieve therapeutic effect range from 40 to 200 mg.

The systemic availability, *F,* of a drug that is absorbed completely from the GI tract after oral administration is the fraction of the absorbed dose that escapes extraction and is given as

$$F = 1 - E \tag{93}$$

Rearranging Equation 85 and substituting for E in Equation 93 yields

$$F = \frac{Q_H}{Q_H + f_{ub}CL_{u,int}} \tag{94}$$

Similar to the limiting conditions described for CL_H, one can define two limiting conditions for systemic availability, F. Specifically, when $Q_H << f_{ub}CL_{u,int}$, F approaches 1.0, whereas when $Q_H << f_{ub}CL_{u,int}$, F approaches zero. These limiting conditions indicate that a drug with a high $f_{ub}CL_{u,int}$ will exhibit low systemic availability after oral administration, because of extensive first-pass metabolism. On the other hand, drugs with a low $f_{ub}CL_{u,int}$ will not be subject to significant first-pass metabolism.

Generally, the parameter most commonly determined to assess overall drug availability after oral administration is the area under the drug concentration-versus-time curve, *AUC*. Recall from Equation 67 that the total clearance is equal to the quotient of intravenous dose and *AUC* from zero to infinity. If the drug is eliminated entirely by metabolism, the hepatic clearance is defined as

$$CL_H = \frac{Dose_{IV}}{AUC_{IV}} = \frac{FDose_{PO}}{AUC_{PO}} \tag{95}$$

Substituting Equation 86 for CL_H and Equation 94 for F yields

$$\frac{(Q_H)(f_{ub}CL_{u,int})}{Q_H + f_{ub}CL_{u,int}} = \left(\frac{Dose_{PO}}{AUC_{PO}}\right)\left(\frac{Q_H}{Q_H + f_{ub}CL_{u,int}}\right) \tag{96}$$

Simplifying,

$$f_{ub}CL_{u,int} = Dose_{PO}/AUC_{PO} \tag{97}$$

Thus, for a high intrinsic hepatic clearance drug, the AUC_{PO} is independent of Q_H. Additionally, the steady-state *AUC* for unbound drug is independent of the free fraction. For a more in-depth discussion of these concepts on hepatic clearance, see the paper by Wilkinson and Shand.[16]

RENAL CLEARANCE—Physiologists studied the renal clearance of endogenous and exogenous substances long before the use of clearance concepts became *popular* in pharmacokinetics. Indeed, the basis for the understanding of drug clearance in pharmacokinetics has its roots in the decades of work by renal physiologists. Moreover, there are significant differences in the complexity of processes involved in hepatic and renal handling of drugs. In the kidney, there are three primary processes (and one minor process) responsible for the renal elimination of drugs, namely filtration, secretion, and reabsorption (and metabolism), respectively. Each of the major processes are affected by common, yet unique, determinants.

The rate of filtration in the kidney for a drug is given as

$$\text{Rate of filtration} = (GFR)(C_u) \tag{98}$$

where *GFR* is the glomerular filtration rate and C_u is the previously defined free-drug concentration. The clearance by filtration is the quotient of the rate of filtration at a given concentration; therefore, the renal clearance, CL_R, of a drug due to filtration is

$$CL_R = (GFR)(f_u) \tag{99}$$

The renal clearance of a drug that is eliminated only by filtration can be estimated if the glomerular filtration rate and the free fraction of drug are known. There are two substances commonly used to estimate the *GFR,* namely creatinine, an endogenous by-product of muscle metabolism, and inulin, a polysaccharide. Both are essentially 100% eliminated in the urine by filtration and, thus, the CL_T of these two substances can be used as reasonable estimates for the *GFR*.

Another primary process involved in the renal elimination of drugs is active tubular secretion, *ATS*. There are several active-transport systems in the proximal renal tubule that are capable of excreting drugs from the flood into the urine. There appears to be a multiplicity of systems for the *ATS* of cations and anions. Whenever the renal clearance is greater than the product of the *GFR* and the free fraction, there must be net tubular secretion in addition to clearance by filtration. The renal clear-

ance due to *ATS,* which is CL_{ATS}, is given as

$$CL_{ATS} = \frac{(Q_{RP})(f_u CL_{u,s,int})}{Q_{RP} + f_u CL_{u,s,int}} \quad (100)$$

where *QRP* and $CL_{us,int}$ represent effective renal plasma flow and unbound intrinsic secretory clearance, respectively. Similar to the situation described for hepatic clearance, drugs may exhibit a high or low intrinsic secretory clearance. The impact of changes in plasma protein–binding on renal clearance would substantially differ between these two situations.

For drugs that undergo both filtration and *ATS,* the renal clearance is simply the sum of the clearance due to filtration and the clearance due to secretion

$$CL_R = (f_u)(GFR) + CL_{ATS} \quad (101)$$

In addition to these two processes, some drugs undergo tubular reabsorption, whereby some fraction of the drug excreted into the urine by filtration and secretion is reabsorbed into the body. Therefore, the full expression for renal clearance, taking into account all three processes, is given by

$$CL_R = (f_u)(GFR) + CL_{ATS} - F_{TR}((f_u)(GFR) + CL_{ATS})$$
$$\text{or } CL_R = (1 - F_{TR})((f_u)(GFR) + CL_{ATS}) \quad (102)$$

where F_{TR} is the fraction undergoing tubular reabsorption.

These relationships provide the basis for determining the primary mechanisms involved in the renal handling of a given drug. Specifically, determination of the ratio of renal clearance and that of inulin provides a clinically useful means to determine the processes that are primarily responsible for renal excretion. If the ratio is equal to 1, the drug would appear to be filtered exclusively by the kidney. Both *ATS* and *TR* also could be occurring, but at equal rates (an unlikely occurrence). If the ratio of renal clearance to inulin clearance is greater than 1, it is clear that the drug is undergoing net tubular secretion as well as filtration. Similarly, if the ratio is less than 1, the drug must be undergoing net tubular reabsorption.

Assessment of the mechanisms for the renal excretion of specific drugs is important because different factors will alter CL_R. For example, if a drug undergoes net tubular secretion, other drugs secreted by the same transport processes may compete for secretory sites, resulting in an overall decrease in the renal excretion of the drug. Alternatively, if a drug undergoes tubular reabsorption and is a weak acid (eg, salicylate) or a weak base (eg, amphetamine), the renal clearance may be altered by manipulation of the urine pH or urine flow. See the review by Tucker[17] for a description of the methods for calculation of CL_R.

PROTEIN-BINDING—Drugs circulating in the blood may bind reversibly to a number of components including plasma proteins. This reversible binding may be described by simple mass-law relationships.

$$[D_{unbound}] + [P] \leftrightarrow [D - P] \quad (103)$$

where $[D_{unbound}]$, $[P]$, and $[D - P]$ are the molar concentrations of the free drug, the protein to which the drug binds, and the drug-protein complex, respectively. It is obvious from this relationship that the amount of drug-protein complex formed is a function both of the concentration of the drug and protein and the affinity between the protein and the drug. Thus, changes in the protein concentration may alter binding, as may changes in the total drug concentration. For most drugs and their respective binding proteins, the concentration of protein far exceeds the concentration of drug, such that the fraction of drug that is bound to protein is independent of drug concentration in plasma or blood.

Because plasma proteins often have a molecular size that restricts their passage across cell membranes and capillary walls, drugs that are bound to plasma proteins are restricted similarly. Thus, plasma protein-binding can have a marked effect on the distribution and elimination of drugs. It is a basic tenet of pharmacology that only the unbound (free) drug is pharmacologically active, because it is assumed that the unbound drug is able to traverse biological membranes and reach the site of drug action. While there have been few direct tests of this hypothesis, those investigations that have been conducted support the assumption that the free drug is the principal pharmacologically active species.

There are several methods by which drug protein-binding may be described quantitatively, though the most frequent and useful is the free fraction, f_u. The f_u can be determined as

$$fu = \frac{C_u}{C_u + C_B} = \frac{C_u}{C_t} \quad (104)$$

where C_u is the concentration of unbound or free drug, C_B is the concentration of drug bound to protein, and C_T is the concentration of total drug (bound plus free). Obviously, f_u can range from 0 to 1. This relationship provides a means by which free drug *in vivo* can be calculated if the total concentration and the free fraction are known.

$$C_u = (C_T)(f_u) \quad (105)$$

Recognizing changes in protein-binding is important because it may substantially alter the pharmacokinetics of a drug. The previous section described the impact of protein-binding on clearance, referenced to total drug concentration. Protein- binding changes also may result in alterations in other pharmacokinetic parameters. The volume of distribution at the steady state can be expressed as

$$V_{ss} = V_{blood} + V_{TW}\frac{f_u}{f_{u_T}} \quad (106)$$

where V_{TW} is the volume of tissue water and f_{uT} is the free fraction of drug in tissue. From the relationship in Equation 106 (which is analogous to Equation 12, for drugs with a volume of distribution >50 L), it is clear that a decrease in protein-binding (ie, an increase in f_u) will result in an increase in the V_{SS}. The impact of changes in f_u on the half-life of a drug with a large volume of distribution can be assessed from Equation 18 and either Equations 87 or 88. The impact resulting from a change in protein-binding of a drug depends upon the magnitude of such binding alterations on both V_{SS} and CL_T.

The two major drug-binding proteins in plasma are albumin and α^1-acid glycoprotein (AAG). Albumin is the major protein both in plasma and in the extracellular space outside of plasma and is present in concentrations ranging from 3.5 to 5.5 g/dL in normal, healthy individuals. Albumin is the primary binding protein for acidic drugs, such as salicylate, tolbutamide, or warfarin. Numerous diseases can result in marked reductions in the concentration of albumin, including nephrotic syndrome, severe burns, liver disease, malnutrition, and some chronic inflammatory conditions.[18] Thus, disease is most likely to produce an increase in f_u for those drugs highly bound to albumin.

The substance AAG belongs to the family of *acute-phase reactants,* endogenous substances that are markedly increased in concentration secondary to some type of stress. While normal AAG concentrations range from 80 to 120 mg/dL, concentrations may increase above 300 mg/dL in patients experiencing major stress, such as surgery, trauma, or burns. More-moderate elevations of AAG have been observed in patients following a myocardial infarction or in inflammatory diseases such as Crohn's disease. The increase in AAG concentration results in a decrease in f_u, and AAG is the major binding protein for many basic, lipophilic drugs.

A major source of drug interactions is competition for protein-binding sites. Each albumin molecule contains at least four different drug-binding sites, two of which are the sites where most drugs that bind to albumin interact with the other molecule. If two drugs bind to the same site on a protein, they may compete with each other for binding. Thus, the addition of a drug to the existing therapeutic regimen of a patient may result in displacement of existing bound drug molecules from their protein-binding sites. However, as described in the *Hepatic Clearance* section of this chapter, these types of interactions rarely are clinically significant (ie, these interactions do

not significantly alter the free-drug concentration). Hence, while protein-binding interactions are probably the most widely reported drug interaction, they rarely necessitate alterations in drug therapy (citation Benet and Hoerner, CPT.

DOSE- AND TIME-DEPENDENT PHARMACOKINETICS

Up to this point, the processes for drug absorption, distribution, metabolism, and elimination have been assumed to be characterized by first-order rate constants, and the general concepts and equations presented are applicable to a wide variety of drugs, with modification. Moreover, with any of the pharmacokinetic models (compartmental, noncompartmental, or physiological), a number of basic assumptions apply, in particular, the principle of superposition holds. In other words, measurements of the concentration of drug plasma, urinary excretion of unchanged drug, or amount of metabolite recovered in bile increase proportionally with increases in dose. When these measurements or other observations are corrected for dose, the values are identical or superimposable. Thus, the pharmacokinetic parameters V_D, CL_T, and F remain constant with respect to time and with dose or concentration.

But the processes controlling the disposition of drugs are biological and therefore involve processes that are mediated by specialized carriers or enzymes. Under some conditions, these processes can become saturated, and changes in dose may produce nonproportional changes (eg, in concentration, amount of metabolite(s) produced, etc.) Table 58-2 delineates some of the various causes of nonlinear pharmacokinetic behavior.

Nonlinearity is a term applied to all situations in which a semilogarithmic plot of plasma concentration versus time data cannot be resolved completely into log-linear components (ie, first-order processes). There are a wide variety of causes for nonlinearity, such as capacity-limited metabolism, capacity-limited absorption, saturable first-pass metabolism, changes in blood supply to the site of absorption and/or the organ of elimination, low or erratic dissolution or release rates from dosage forms, low solubility of the drug, or drug-induced changes in organ function or body temperature. Nonlinear drug disposition primarily has been determined by measuring the pharmacokinetics at several dosage levels. When a capacity-limited enzyme metabolism is the source of the nonlinearity, the Henri-Michaelis-Menten equation

$$Velocity = \frac{V_{MAX}C_{ss}}{K_M + C_{ss}} \qquad (107)$$

can be applied to assess the velocity versus substrate (drug) concentration relationship. There are several techniques for the determination of the direct cause of nonlinearity in drug kinetics, including direct calculation of CL_T, CL_{ORAL}, F, V_{SS}, and V_1. Most commonly, lack of superposition (disproportionate increase in AUC with increasing dose) is an indication of nonlinearity in the system.

PROTEIN-BINDING—When the amount of drug bound to plasma proteins approaches saturation, the percentage of the drug that is unbound may vary considerably with increasing dose. Under the conditions of saturation, for example, after a salicylate overdose, the f_u may vary considerably with the total amount of drug in the body and hence, certain pharmacokinetic parameters, such as apparent volume of distribution, will be influenced.

TIME-DEPENDENT KINETICS—Carbamazepine is a drug with low-to-intermediate intrinsic clearance, which also induces an increase in the activity of the biotransforming enzyme system by which it is metabolized. This increase will increase total clearance and decrease half-life. Because such autoinduction of metabolism does not occur until several dose-intervals of repetitive dosing, the pharmacokinetics vary with time and are called time-dependent. *Allosteric* (feedback) inhibition by accumulated metabolites of a drug, or an effect of a drug to impair its route of elimination, also will cause dose- and time-dependent changes in pharmacokinetics. Drugs that cause the depletion of some slowly replaceable intermediary factor, such as the depletion of norepinephrine by reserpine or the depletion of inorganic sulfate by acetaminophen, will manifest time-dependent effects.

Table 58-2. Examples of Mechanisms for Dose- and Time-Dependent Pharmacokinetics (Nonlinear Drug Disposition)

KINETIC PROCESS AND MECHANISM	EXAMPLES
Gastrointestinal absorption	
Saturable transport	Riboflavin, penicillins
Intestinal metabolism	Salicylamide
Biotransformation	
Saturable metabolism	Phenytoin, salicylate
Product inhibition	Phenytoin (rat)
Cosubstrate depletion	Acetaminophen
Plasma protein-binding	Prednisolone, disopyramide
Renal excretion	
Glomerular filtration/ protein-binding	Naproxen
Tubular secretion	*p*-Aminohippuric acid, mezlocillin
Tubular reabsorption	Riboflavin, cephapirin
Biliary excretion	
Biliary secretion	Iodipamide, BSP
Enterohepatic cycling	Cimetidine, isotretinoin
Tissue distribution	
Plasma protein-binding	Prednisolone, ceftriaxone
Hepatic uptake	Indocyanine green, warfarin (rat)
CSF transport	Benzylpenicillins
Cellular uptake	Methicillin (rabbit)
Tissue-binding	Cyclosporine, dideoxyinosine (rat)

DOSE-DEPENDENT KINETICS—When the elimination route is saturated by either capacity-limited metabolism or capacity-limited renal excretion, it is evident that the total clearance of a drug will decrease and the half-life will increase with increases in doses. Examples of important drugs that demonstrate dose-dependent kinetics are salicylic acid, phenylbutazone, probenecid, levodopa, phenytoin, heparin, and dicumarol. Ethanol obeys essentially zero-order elimination kinetics at blood concentrations above 0.02 to 0.04%, which is a fact of considerable social and legal importance.

Salicylic acid is one of the most interesting examples of dose-dependent kinetics from multiple sources. Salicylic acid is eliminated from the body by at least five, parallel, competing processes.[19] Two of these are saturable processes for the formation of salicyluric acid (the glycine conjugate of salicylic acid) and salicylphenol glucuronide. The other three processes of elimination, excretion of unchanged salicylic acid in urine and formation of gentisic acid and salicyl glucuronide are first-order processes. The half-life of salicylic acid increases from about 3 hr to over 20 hr as the dose is increased upward from 300 mg to 10 g. At low doses, the half-life is about 3 hr, the apparent volume of distribution is approximately 9L, and a total clearance can be estimated to be 2 L/hr. The binding of salicylic acid to albumin also is capacity-limited (saturable), and saturation occurs even at therapeutic (low) doses of the drug. Therefore, as the amount of salicylic acid in the body increases, the $CL_{u,int}$ decreases, whereas the f_u increases. These two effects tend to oppose each other, such that the total clearance of salicylic acid remains relatively unchanged within the anti-inflammatory range of unbound concentrations (between 10 and 60 mg/L; Fig 58-19).[20] Finally, the renal excretion of salicylic acid (pK_a = 3.5) can be increased by increasing urinary pH, resulting in a decrease in renal tubular reabsorption (data not shown). The toxicological consequences of a salicylic acid overdose are well known, but it is not always appreciated that as the concentration of salicylic acid increases, the total amount of drug in the body increases out of proportion to the total plasma concentration.

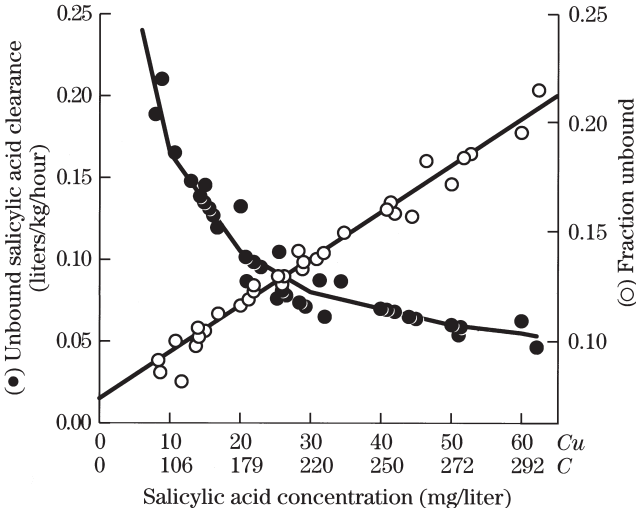


Figure 58-19. The clearance of unbound drug (●), determined under steady-state conditions, and the fraction unbound in plasma (○) vary inversely with each other as the salicylic acid concentration is increased. The corresponding total plasma salicylic acid concentrations are superimposed on the linear scale of the concentration of unbound drug; 1 mg/L = 7.2 micromolar. (Redrawn from Furst DE, Tozer TN, Melmon KL. *Clin Pharmacol Ther* 1979; 26:380.)

STEREOCHEMICAL CONSIDERATIONS

Chiral drugs constitute approximately 60% of the drugs that are currently commercially available. Most of these are marketed as racemic mixtures. These facts obviously indicate the importance of understanding the impact of stereochemistry on both pharmacokinetics and pharmacodynamics. While it has long been appreciated that optical isomers often differ in the potency of pharmacological or toxic effect, it is only recently that significant attention has been paid to the influence of chirality on pharmacokinetic processes involved in absorption, distribution, and elimination. This is primarily due to the previous lack of analytical methodology required to separate drug enantiomers. The recent development of reasonably inexpensive methods for the separation of stereoisomers has led to a more comprehensive assessment of the pharmacokinetics of drugs that are administered as racemic mixtures.

Enantiomers possess identical physical and chemical properties, despite significant differences in spatial configuration. Thus, biological processes that are passive in nature (and thereby depend only upon physical and chemical characteristics of the molecule) do not display selectivity for one isomer over another. In contrast, biological processes that require the interaction of a drug molecule with a macromolecule (such as protein-binding or metabolism) may exhibit stereoselectivity. This knowledge permits some generalizations about when pharmacokinetic processes may differ between enantiomers.

ABSORPTION—Since most drugs are absorbed by passive diffusion, most will not exhibit stereoselective alterations in absorption. Drugs that are absorbed by a carrier-mediated or active process may display such stereoselectivity. Indeed, demonstration of stereoselective absorption would be strong evidence that a drug is absorbed via a carrier-mediated process.

PROTEIN-BINDING—Drug association with plasma proteins requires interaction of a small molecule with a macromolecule, which depends upon the spatial configuration of both components. It should not be surprising, therefore, that plasma protein-binding has been found to exhibit stereoselectivity for some drugs, including disopyramide, ibuprofen, mexilitene, propranolol, and verapamil.

METABOLISM—Biotransformation requires the interaction of a drug with an enzyme, an interaction in which spatial arrangement is critical. Many drugs that undergo metabolism exhibit stereoselective hepatic clearance. For example, the oral clearance of verapamil (a high intrinsic clearance drug) displays profound stereoselectively such that the oral clearance ratio of the *R* to *S* isomers is approximately 4.

RENAL EXCRETION—Filtration in the kidney is a passive process; however, if a drug exhibits stereoselective protein-binding, one might anticipate that the drug enantiomers would exhibit differential filtration rates. Active tubular secretion, being an active process, also may demonstrate stereoselectively for some drugs. Indeed, numerous drugs, including chloroquine, disopyramide, and terbutaline, have been found to be secreted stereoselectively by the kidney. While passive tubular reabsorption would not be expected to show stereoselective effects, active reabsorption may demonstrate these effects as has been shown for certain endogenous substances such as glucose and amino acids.

KINETICS OF PHARMACOLOGIC EFFECT

In addition to considering the relationship between drug concentration and time, proper design of therapeutic regimens often necessitates an understanding of the relationship between concentration and response. These two relationships can be combined to produce a time course of pharmacologic response, and this is referred to as pharmacodynamics. This area of investigation provides important quantitative information regarding the onset, duration, and intensity of pharmacologic effect often in relation to drug concentration. Full characterization of these elements can involve the development of pharmacokinetic-pharmacodynamic models. Modeling the kinetics of effect requires an understanding of the mechanism of pharmacologic action, of which there are many. These include receptor stimulation (eg, β_2-adrenoceptor agonists), receptor antagonism (eg, H_1-histamine receptor antagonists), transporter inhibition (probenecid, diuretics), enzyme inhibitors (eg, angiotensin converting enzyme), substrate replacement (eg, thyroxine, testosterone), non-receptor-mediated drug action (eg, chelation) and chemotherapy (eg, antibiotics).

Clinically useful information can often be derived from relatively simplistic models. The relationship between drug response and concentration is usually graded, that is, the rise in concentration results in a progressively increasing magnitude of effect (Fig 58-20). There is essentially always a ceiling, or maximum, to the intensity of effect, such that further increases in drug concentration will not result in additional increases in the intensity of pharmacologic response. In practice, plasma concentrations rarely, if ever, reach the maximum effect because toxicities often develop. Thus, only the linear portion of the concentration-effect curve may be observed. The placement of the curve along the *x*-axis may vary among compounds, resulting in differences in the potency the drugs. Moreover, the steepness of the curve (the rate at which the intensity of the effect changes as a function of concentration) will vary among drug responses. For some drugs, such as narcotic agonists, a small change in the drug concentration can result in marked changes in drug effect (see Fig 58-20).

There are certain agents that appear to exhibit an all-or-none response. Rather than the typical graded increase in effect as concentration increases, these compounds exhibit a threshold of response that once reached result in the detection in the response. While a given individual's response to the drug is identified by the presence or absence of the effect, the fraction of subjects in a sample population responding will increase as the concentration of drug is increased. A good example of this is seen in the conversion to normal sinus rhythm in patients with atrial fibrillation treated with an antiarrhythmic agent (Fig 58-21). In this instance, there is variation within the patient population in the concentration that must be achieved to convert a specific patient from atrial fibrillation to normal sinus rhythm. Thus, the higher the concentration achieved in a patient, the more likely they are to exhibit the pharmacologic effect.

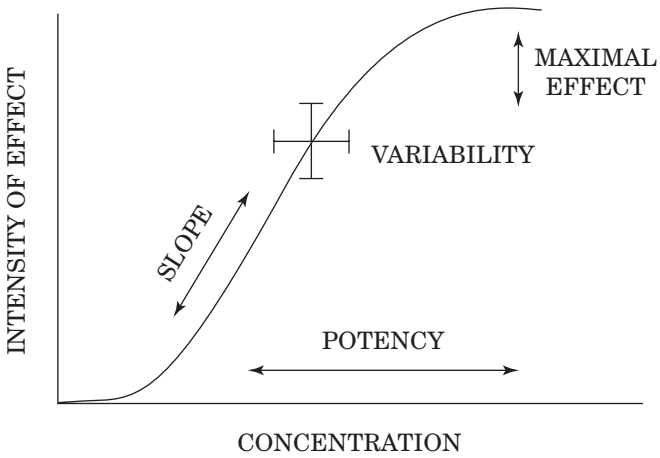


Figure 58-20. A graded pharmacologic response to a drug, showing a progressive increase in the intensity of effect as concentration increases, until the maximal effect is achieved when effect is plotted as a function of the log concentration. (Redrawn from Nies AS, Spielberg SP, Principles of Therapeutics, In: *Goodman and Gilman's The Pharmacologic Basis of Therapeutics*, 9th edition, Hardman JG, Limbird LE, Molinoff PB, Ruddon RW (eds). McGraw Hill, San Francisco, 1996.)

Although visual inspection of concentration-response graphs can yield much useful information, rigorous comparisons between agents requires a more quantitative analysis of concentration response data. The models that are used for such analyses are highly dependent upon the mechanism of action of the drug. In general, the mechanisms of pharmacologic agents can be divided into three categories: direct acting reversible agents, indirect acting reversible agents, and irreversibly acting agents.

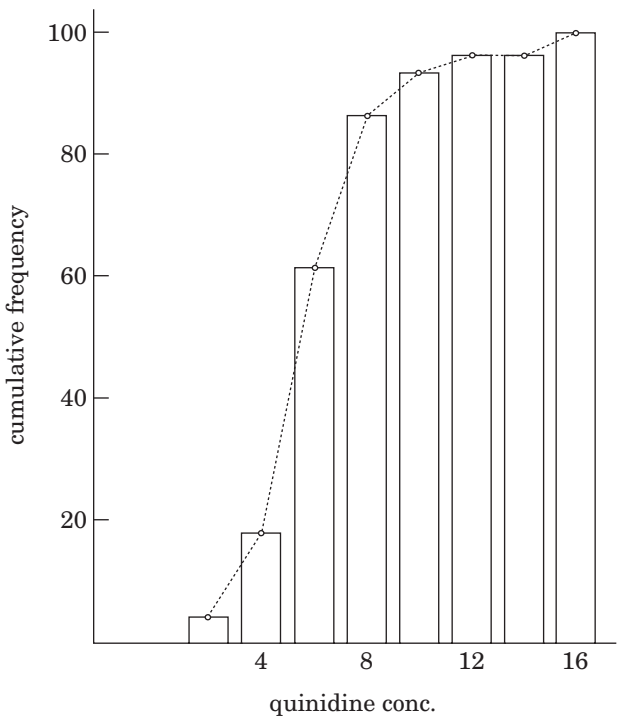


Figure 58-21. Cumulative frequency of conversion to normal sinus rhythm (expressed as percent of patients) as a function of quinidine concentration in a group of subjects with atrial fibrillation treated with the antiarrhythmic quinidine. (Redrawn from Gibaldi M, *Biopharmaceutics and Clinical Pharmacokinetics*, 4th edition, Lea & Febiger, 1991.)

INTENSITY OF EFFECT—The intensity (or magnitude) of pharmacologic effect is given as

$$Intensity\ of\ Effect = \frac{E_{\max} \times C^{\gamma}}{EC_{50}{}^{\gamma} + C^{\gamma}} \qquad (108)$$

where E_{max} = maximum effect, EC_{50} = concentration necessary to achieve 50% E_{max},

C = concentration of drug, and γ = Hill coefficient.$_x$

This equation is similar to that which describes the binding of oxygen to hemoglobin. One significant difference is that when used to quantify pharmacologic effect, the Hill coefficient has no physical or mechanistic meaning. However, the larger the value of the Hill coefficient, the steeper will be the concentration-response relationship. For some measured pharmacologic responses, there is a baseline physiological value that must be taken into consideration when modeling the pharmacologic effect. For example, there are a number of therapeutic agents that induce methemoglobinemia after ingestion. There is, however, a low level of methemoglobin (1-2%) in drug-naïve subjects that needs to be accounted for in modeling the intensity of effect of agents that induce methemoglobinemia. Equation 108 can be adjusted to account for baseline effect:

$$Intensity\ of\ Effect = \frac{E_0 + E_{\max} \times C^{\gamma}}{EC_{50}{}^{\gamma} + C^{\gamma}} \qquad (109)$$

where E_0 = baseline effect.

Such a quantitative approach permits comparison of the intensity of drug effect between various agents. For example, while the antimicrobial agent dapsone is widely reported to cause methemoglobinemia in patients treated with the drug, the antimicrobial agent sulfamethoxazole is rarely associated with methemoglobinemia; despite the fact that both form a reactive hydroxylamine metabolite and that sulfamethoxazole is administered at much higher doses than dapsone. It was unclear whether this difference was due to differences in the potency of their respective hydroxylamine metabolite or due differences in the pharmacokinetics of the drugs and/or metabolites. This question was addressed by comparing the ability of the two hydroxylamine metabolites for forming methemogobin when incubated with human erythrocytes *in vitro* and modeling the effects using Equation 109.[21] As shown in Table 58-3, the maximal effect of the two metabolites studied were similar, but the potency (EC_{50}) of dapsone hydroxylamine was 20-fold greater than that of sulfamethoxazole hydroxylamine.

DURATION OF EFFECT—A measurable pharmacologic effect will be observable as soon as drug concentration reaches the minimal effect concentration (MEC). The duration of effect is determined by how long the concentration remains above the MEC, which is influenced by the dose administered and the rate of elimination of the drug. Consider the linear range of the log concentration-effect curve, such that $0.2E_{max} < E < 0.8E_{max}$. The effect within this boundary may be expressed as

$$E = m\log C + r \qquad (110)$$

Table 58-3. Pharmacodynamics of Dapsone Hydroxylamine (DNOH) and Sulfamethoxazole Hydroxylamine (SNOH) Induced Methemoglobin Formation in Human Erythrocytes

	DNOH	SNOH
E_{max} (%)	26 (5)	463 (105)
EC_{50} (μM)	89 (4)	84 (5)

Data from Reilly TP, et al. *J Pharmacol Exp Ther* 199; 288:951.

where m = slope of E versus log C plot and r = constant. This equation can be rearranged, such that

$$\log C = \frac{E - r}{m} \tag{111}$$

Recall that for a drug exhibiting instantaneous distribution, following instantaneous administration

$$\log C = \log C_0 - \frac{\lambda}{2.303} t \tag{112}$$

The maximal effect elicited by this dose, E_m, will occur when $C = C_0$. Thus,

$$\log C_0 = \frac{E_m - r}{m} \tag{113}$$

Substituting the equivalents in Eq. 111 and 113 into Eq. 112 yields

$$E = E_m = \frac{m\lambda}{2.303} t \tag{114}$$

Thus, the intensity of effect should decline at a constant rate that is a function of the elimination rate constant and the slope of the response versus log concentration curve. It should be noted that though the decrease in concentration is first-order, the decrease in effect is zero-order. But how can we quantify the actual duration of action following a given dose of a drug? Recall from Equation 13 that

$$C = C_0 e^{-\lambda t}$$

Expressed in terms of dose

$$C = \frac{D_{iv}}{V_d} e^{-\lambda t} \tag{115}$$

The duration time (t_d) is that time at which the plasma concentration drops just below the *MEC*, such that

$$MEC = \frac{D_{iv}}{V_d} e^{-\lambda t} \tag{116}$$

Solving to duration time

$$t_d = \frac{1}{\lambda} [\log D_{iv} - \log (MEC) V_d] \tag{117}$$

DIRECT ACTING, REVERSIBLE AGENTS—It could be argued that few, if any, drugs are truly 'direct acting.' Most drugs interact with a receptor that produces the effect. Sometimes this interaction results in a cascade of events that ultimately produce the measured pharmacologic response. Thus, it might be appropriate to designate a category of drugs as *rapid acting, reversible agents.* The drug-receptor interaction is easily reversible, the pharmacological effect is easily reversible, and the time course of the effect is not delayed with respect to the time course of the drug-receptor interaction. For example, the β_1-adrenoceptor antagonists (metoprolol) and the suppression of exercise-induced heart rate. There is a reversible (ie, non-covalent) interaction with a cellular macromolecule, which, as a consequence of the binding of the drug, stimulates a cellular response. This can expressed as

$$C \leftrightarrow C_R + R \leftrightarrow C_R - R \Rightarrow RESPONSE$$

where C = concentration in plasma, C_R = concentration at receptor site, R = receptor, $C_R - R$ = drug receptor complex.

The intensity and duration of response can be readily quantified by measuring the 'response' at several different concentrations.

INDIRECT ACTING, REVERSIBLE AGENTS—For many drugs, the response measured clinically is several steps removed from the initial biochemical effect of the drug. The time course of the effect therefore lags behind the time course of concentrations. In these circumstances, there may appear to be no direct association or relationship between the concentration of the drug in blood or plasma and the pharmacologic response.

$$C \leftrightarrow C_R + R \leftrightarrow C_R - R \Rightarrow (\Rightarrow \Rightarrow \Rightarrow) RESPONSE$$

and "⇒" represents a 'transduction' of the response that may depend on several factors including the rate of turnover and trafficking of endogenous substrates or other mediators of drug effect. In this instance, the drug-receptor interaction may not easily be reversible and the pharmacologic effect is prolonged, such as the acetylation by aspirin of the serine moiety at the active site of cyclooxygenase. For some drugs with an indirect pharmacologic response, the drug-receptor interaction is easily reversible but the pharmacologic effect is prolonged. An example might be corticosteroids which bind with nuclear receptor proteins resulting in RNA transcription and protein synthesis.

A classic example of an indirect pharmacologic response is the oral anticoagulant warfarin (Fig. 58-22). In this instance, the dissociation between the pharmacologic effect and the concentration is readily understood by a consideration of the mechanism action. Warfarin inhibits the synthesis of the vitamin K-dependent factors that determine the prothrombin time. However, there is a time lag for this effect to be observed since circulating clotting factors must be degrading by the normal metabolic processes before an anticoagulant effect is observed. Thus, the turnover rate of the clotting factors and not the time course of warfarin concentrations, becomes key determinant for the onset, duration and offset of effect. This was accounted for by models that incorporated the endogenous degradation and synthesis rates of the clotting factors, thereby providing a good relationship between the concentration and 'true' pharmacologic effect.[22]

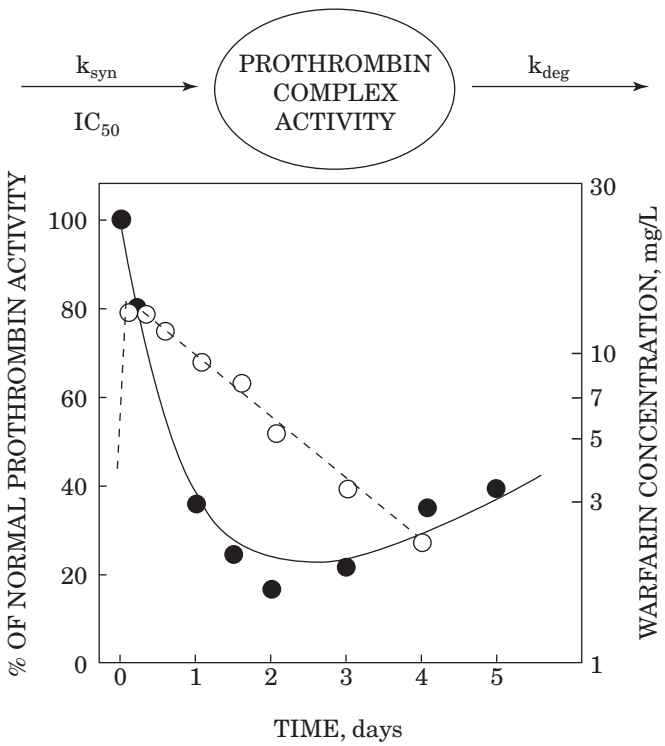


Figure 58-22. Time course of plasma warfarin concentration (top panel) and pharmacologic response, measured as % decrease in normal prothrombin activity (bottom panel). Note that the pharmacologic effect does not become evident until the concentration has reduced substantially from its maximal value. (Redrawn from Nagashima R, et al. *Clin Pharmacol Ther* 1969; 10:22.)

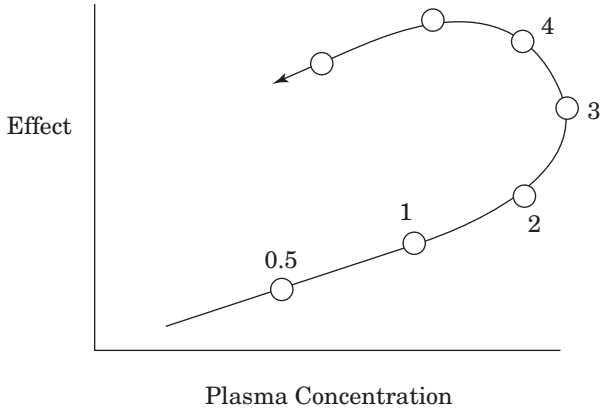


Figure 58-23. A counterclockwise hysteresis is observed following administration of a drug that exhibits non-instantaneous distribution and for which the site of action is in the slowly equilibrating tissues. Open circles represent a single effect measurement at a given concentration, and numbers indicate time (in hours) that measurements were made.

If distribution of a drug to its site of action is delayed, a counterclockwise hysteresis curve can be observed when plotting the effect versus plasma concentration after a dose of the drug, particularly when administered via a slow infusion (Fig 58-23). Thus, the intensity of effect observed for a given concentration shortly after administration (while concentrations are rising) is less than the intensity observed at that same concentration during the decline in concentration. This is due to the fact that during the distribution phase most of the drug is in the rapidly equilibrating tissues (ie, the central compartment), while in the latter time periods the larger fraction of the drug is in the slowly equilibrating tissues (ie, the peripheral compartment). Hence, the relationship between effect and *plasma* concentration for such a drug will be time dependent. This type of hysteresis can also be observed if the response is due to an active metabolite rather than the parent drug. A clockwise hysteresis is observed when tolerance occurs.

Sheiner et al (*CPT* 25:358, 1979) and Sheiner and Holford (*Clin PK* 6:429, 1982) originally modeled the time delay in pharmacological response with a hypothetical 'effect compartment.' Using regression analysis to estimate the rate constants for the 'effect compartment,' the hysteresis loop was collapsed, thereby making it possible to model indirect effects. Jusko and colleagues[24] have developed a series of comprehensive physiologic indirect response models that incorporate an understanding of the mechanism(s) producing a particular pharmacologic effect. These indirect response models envision a given pharmacologic response that is based on the inhibition or stimulation of the input or the output factors controlling the pharmacologic effect. These complex models have increased our understanding of a variety of pharmacologic effects, such as induction of protein synthesis, cell trafficking, altered hormone secretion.[23]

IRREVERSIBLE ACTING AGENTS—Modeling the pharmacodynamics of irreversible acting agents is considerably more complex, but has been successfully accomplished. The most intensely studied agents are those used in the treatment of cancer or infections. Modeling the ability of such agents to kill tumor or bacterial cells necessitates incorporation of cell-cycle kinetics. Such models have been useful in determining whether agents are best administered infrequently at high doses or in continuous exposure regimens.

REFERENCES

1. Swerdloff RS, et al. *Diabetes* 1967; 16:161.
2. Edwards DJ, et al. *Clin Pharmacokinet* 1982; 7:421.
3. Sulh H, et al. *Clin Pharmacol Ther* 1986; 40:604.
4. Barr WH. *Am J Pharm Educ* 1968; 52:958.
5. Øie S, Tozer TN. *J Pharm Sci* 1979; 68:1203.
6. Williams RL, et al. *Clin Pharmacol Ther* 1977; 21:301.
7. Truitt EB Jr, et al. *J Pharmacol Exp Ther* 1950; 100:309.
8. Gibaldi M, Perrier D. *Pharmacokinetics,* 2nd ed. New York: Dekker, 1982, pp 30, 276, 441.
9. Jusko WJ. In *Applied Pharmacokinetics:. Principles of Therapeutic Drug Monitoring,* 3rd ed. Evans WE, Schentag JJ, Jusko WJ, eds. Spokane, WA: Applied Therapeutics, 1992, 2-1.
10. Nuesch EA. *Drug Metab Rev* 1984; 15:103.
11. Yeh KC, Kwan KC. *J Pharmacokinet Biopharm* 1978; 6:79.
12. Benet LZ, Galeazzi RL. *J Pharm Sci* 1979; 68:1071.
13. Karol MD. *Biopharm Drug Dispos* 1990; 11:179.
14. Bigger JT. *Am J Med* 1975; 58:479.
15. Morgan DJ, Smallwood RA. *Clin Pharmacokinet* 1990; 18:61.
16. Wilkinson GR, Shand DG. *Clin Pharmacol Ther* 1975; 18:377.
17. Tucker GT. *Br J Clin Pharmacol* 1981; 12:761.
18. Svensson CK, et al. *Clin Pharmacokinet* 1986; 11:450.
19. Levy G, Tsuchuya T. *N Engl J Med* 1972; 287:430.
20. Furst DE, Tozer TN, Melmon KL. *Clin Pharmacol Ther* 1979; 26:380.
21. Reilly TP, et al. *J Pharmacol Exp Ther* 1999; 288:951.
22. Nagashima R, et al. *Clin Pharmacol Ther* 1969; 10:22.
23. Sharma A, Jusko WJ. *Br J Clin Pharmacol* 1998; 45:229.
24. Jusko WJ, Ko HC. *Clin Pharmacol Ther* 1994; 56:406.

BIBLIOGRAPHY

Evans WE, Schentag JJ, Jusko WJ. *Applied Pharmacokinetics. Principles of Therapeutic Drug Monitoring,* 3rd ed. Spokane, WA: Applied Therapeutics, 1992.

Gibaldi M. *Biopharmaceutics and Clinical Pharmacokinetics,* 4th ed. Philadelphia: Lea & Febiger, 1991.

Gibaldi M, Perrier D. *Pharmacokinetics,* 2nd ed. New York: Dekker, 1982.

Pecile A, Rescigno A. *Pharmacokinetics. Mathematical and Statistical Approaches to Metabolism and Distribution of Chemicals and Drugs.* New York: Plenum Press, 1988.

Pratt WB, Taylor P. *Principles of Drug Action. The Basis of Pharmacology,* 3rd ed. New York: Churchill Livingstone, 1990.

Reidenberg MM, Erill S, eds. *Drug-Protein Binding,* Esteve Found Symp I. New York: Praeger, 1986.

Rowland M, Tozer TN: *Clinical Pharmacokinetics. Concepts and Applications,* 3rd ed. Philadelphia: Lea & Febiger, 1995.

Shargel L, Yu ABC. *Applied Biopharmaceutics and Pharmacokinetics*, 4th ed. Norwalk, CT: Appleton & Lange, 1999.

Winter ME. *Basic Clinical Pharmacokinetics,* 3rd ed. Spokane, WA: Applied Therapeutics, 1994.

CHAPTER 59

Clinical Pharmacokinetics and Pharmacodynamics

Paul M Beringer, PharmD
Michael E Winter, PharmD

In Chapter 58, the basic principles of pharmacokinetics were presented. Clinical pharmacokinetics is the discipline in which basic pharmacokinetic principles are applied to the development of rational dosage regimens. In this chapter, the concepts of pharmacokinetics are placed into perspective with the development of individualized drug dosage regimens. The clinical significance of drug absorption, distribution, and elimination and influence of disease states on these processes are emphasized. Examples are given of the ways pharmacokinetic principles can be applied in the calculation and adjustment of dosage regimens designed to fit the pharmacokinetic and pharmacodynamic properties of drugs and specific disease states that alter drug disposition. The principles of therapeutic drug monitoring and the rational use of this clinical science in the management of patients also are discussed.

Overview of Clinical Pharmacokinetics

The application of pharmacokinetic principles to patient care can aid the clinician in making rational drug use decisions. However, knowing the relationship between the time course of drug concentration and the pharmacologic effect is critical to the application of pharmacokinetic principles and the interpretation of plasma drug concentrations in the patient care setting.

As a general rule traditional pharmacokinetic research is an intensive study of a limited number of subjects resulting in very precise pharmacokinetic and pharmacodynamic parameter estimates. Clinical pharmacokinetics, on the other hand, is usually limited to very few and sometimes no plasma drug concentrations, requiring the clinician to make an educated guess about key elements of drug disposition and the drug use process. In the research setting, it is common to obtain 10 or more samples for drug concentration measurements within a single dosing interval. In the clinical setting, it is uncommon to obtain more than two or three samples for a patient during a hospitalization or within a year for ambulatory care patients.

Therefore, understanding the usual manner in which drugs are absorbed, distributed, and eliminated as well as the known factors that alter drug disposition and which of these elements is most likely to be altered in the individual patient is key to the clinician's ability to effectively use pharmacokinetics. A basic knowledge of pharmacokinetics provides guidance to the clinician when selecting a drug product, dosing regimen, the anticipated onset of drug effect, and determining an appropriate sampling strategy if drug concentrations are to be obtained.

Drugs with Narrow Versus Wide Therapeutic Range

The therapeutic range is a concentration range that is likely to result in the desired clinical or therapeutic response with an acceptable risk or likelihood of developing a toxic response. For every drug, there is a therapeutic range, but it is those drugs in which the minimum concentration that is likely to result in the desired drug effect is relatively close to the higher drug concentration that is likely to result in a toxic response. The therapeutic index is the ratio of the maximum desired concentration relative to the minimum desired concentration. The application of pharmacokinetic principles may be limited in the use of some drugs. Drugs that have a wide therapeutic index may not require precise dose adjustments when drug disposition is altered and a simple approximation may be satisfactory to limit the probability of toxicity and assure efficacy. Other drugs may have a complex series of biological events that result in an obscure relationship between the pharmacologic effects and the drug dose or drug concentration making it difficult to apply the usual pharmacokinetic principles to the daily care of a patient.

Drugs with a narrow therapeutic range, however, tend to lend themselves to careful dose adjustments and plasma drug concentration monitoring to help ensure optimal patient outcomes. For those drugs that are monitored with plasma drug concentrations, there is usually a *Normal Therapeutic Range* that attempts to define the drug concentrations where the benefit to risk ratio is optimal (Fig 59-1). While the Normal Therapeutic Range is important, it is only a guide, and it is the patient and not the drug concentration that is therapeutic or toxic. There are patients with an optimal clinical outcome whose plasma drug concentrations fall outside the usual range and others who develop unacceptable side effects or toxicities when drug concentrations are within or even below the usual Therapeutic Range.

Plasma Protein Binding and the Therapeutic Range

One potential factor that can change the Normal Therapeutic Range is alterations in plasma protein binding. In most cases, clinical laboratories use assay procedures that measure and report the total plasma drug concentration, ie, the drug concentration that is bound to plasma protein and the unbound plasma drug concentration. It is only the unbound drug in plasma that can cross into the tissue where the receptors are located.

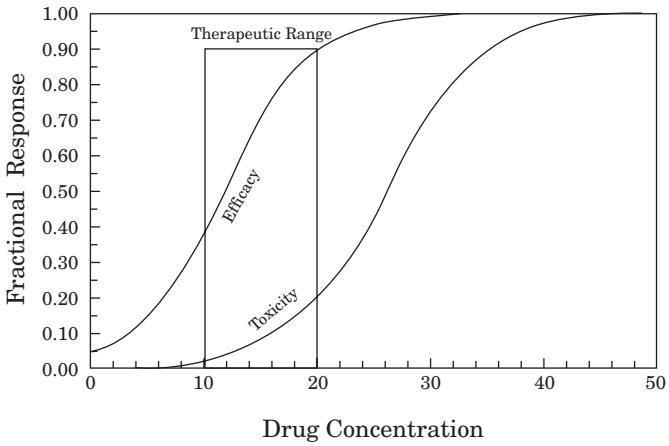


Figure 59-1. "Normal therapeutic range." The "normal therapeutic range" defines the region of drug concentrations where the probability of a positive therapeutic response is good and the risk for development of a significant dose-related adverse effect is acceptable. For most agents the normal therapeutic range is quite wide; however, for certain agents there is a relatively narrow therapeutic range and monitoring of drug concentrations may be necessary to maximize the potential for efficacy and minimize the risk of toxicity.

Therefore, it is the unbound drug concentration that is proportional to the tissue and receptor drug concentration and the pharmacodynamic response (Fig 59-2). Any change in plasma protein binding would be expected to alter the potential for any plasma drug concentration, reported as both bound and unbound drug, to result in a toxic or therapeutic response. Many drugs have significant binding to plasma proteins and the relationship between the unbound drug concentration and the total drug concentration is referred to as the free fraction or fu.

$$\text{fu} = \frac{\text{Unbound Drug Concentration}}{\text{Total Drug Concentration}} \quad (1)$$

or

$$\text{Unbound Drug Concentration} = (\text{fu})(\text{Total Drug Concentration}) \quad (2)$$

Any factor that alters plasma protein binding will result in an altered free fraction (fu'). Therefore, when interpreting assayed drug concentrations with altered plasma binding, the clinician should make some type of adjustment when using the assayed drug concentration.

One approach is to calculate a normal plasma binding drug concentration:

$$\text{Normal Plasma Binding Drug Concentration} = \left(\frac{\text{fu}'}{\text{fu}}\right)\left(\begin{array}{c}\text{Assayed Drug Concentration}\\ \text{with}\\ \text{Altered Plasma Binding}\end{array}\right) \quad (3)$$

and then compare the normal plasma binding drug concentration to the normal therapeutic range to evaluate the drug's potential for either efficacy or toxicity

An alternative approach is to calculate an adjusted therapeutic range:

$$\text{Adjusted Therapeutic Range} = \left(\frac{\text{fu}}{\text{fu}'}\right)(\text{Normal Therapeutic Range}) \quad (4)$$

and compare this adjusted therapeutic range to the assayed drug concentration with altered plasma binding to evaluate the drug's potential for either efficacy or toxicity.

In any case for drugs with high plasma binding, care should be taken in the interpretation of assayed drug concentration. Most weak acid drugs with high plasma binding (eg, phenytoin) are bound almost exclusively to albumin. The most commonly encountered reasons for alterations in plasma binding for these drugs are hypoalbuminemia, end stage renal failure or dialysis and displacement by other drugs. Basic compounds tend to have a more complex plasma-binding pattern and extensive binding to a number of plasma proteins including alpha-1-acid-glycoprotein, other globulins and albumin is common.

In addition to plasma binding alterations, clinical conditions can change a patient's response to a given dose or drug concentration. As an example, a change in renal function can change a patient's ability to eliminate drugs whose route of elimination is via the kidneys (eg, aminoglycoside antibiotics). The addition of a new drug that either inhibits the elimination or metabolism (eg, amiodarone when added to a patient receiving digoxin) or induces metabolism (eg, carbamazepine inducing the metabolism of warfarin) can alter the relationship between the drug dosing regimen and the resultant drug concentrations and drug effect. In addition, alterations in electrolyte or acid base balance might alter the potential of a drug to produce toxicity (eg, hypokalemia in a patient receiving digoxin). While the normal therapeutic range is usually thought of as fixed upper and lower boundaries, there are many situations that require the clinician to make adjustments in dosing regimens and target drug concentrations. Knowing both the drug's pharmacokinetic and pharmacodynamic characteristics allow the clinician to design drug regimens that have an optimal chance of producing a beneficial outcome for the patient.

Absorption, Distribution, and Elimination

In the application of pharmacokinetics to the clinical practice setting, the ability to estimate a patient's absorption, distribution, and elimination characteristics is an important step in initiating drug therapy. For many drugs, clinicians simply learn the "usual" dose and use that dose for all patients. In a number of situations, knowing the principles behind the usual dose allows the clinician to make adjustments in drug therapy for those patients where therapeutic problems, toxicity or lack of efficacy, are likely to occur.

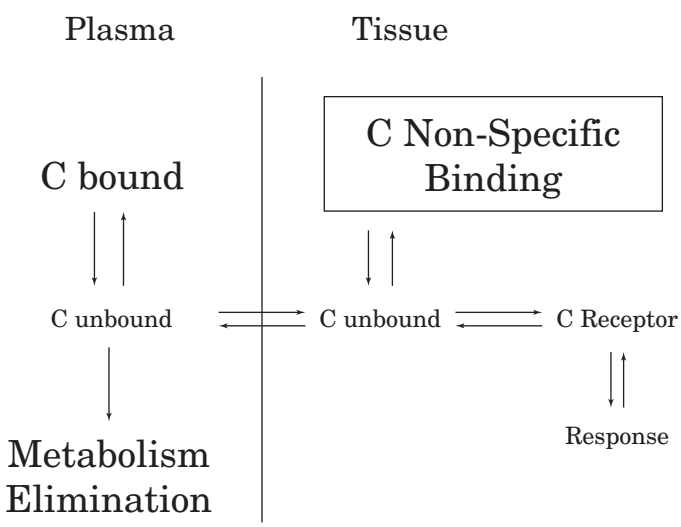


Figure 59-2. Note that it is the unbound drug concentration (C unbound) that is able to cross into the tissue and equilibrate with the tissue binding sites and the drug receptor. While C bound may be a significant percentage of the plasma drug pool, in most cases very little drug is in plasma and therefore C bound represents relatively little of the total drug in the body.

Absorption/Bioavailability

The absorption of a drug is a key element in determining a drug-dosing regimen. The extent of absorption is referred to as bioavailability and is usually expressed as either a fraction (F) or percent of an administered drug that is available to produce a pharmacologic effect. An F value of 1 represent 100% of an administered dose is bioavailable. Most drugs when given by the intravenous route are assumed to be 100% bioavailable (F of 1.0). Absorption by other routes of administration (oral, rectal, etc) may or may not be complete. A number of factors influence the bioavailability of a drug.

To be orally absorbed, drugs must have a reasonable degree of water solubility so that they can dissolve in the gastrointestinal (GI) fluids. In addition, they must also have some lipid solubility characteristics so that the drug can cross the lipid membranes of the cell wall in the GI tract and enter the general circulation and eventually cross the cell walls of other tissues in the body. Aminoglycoside antibiotics are an example of a drug class whose water solubility is so high (lipid solubility very low) that they are not absorbed to any significant extent when administered by the oral route and must be given parenterally to achieve systemic effects.

Drugs that are unstable in the GI tract may have low bioavailability because they are broken down or decompose before they can be absorbed. The proton pump inhibitors (eg, omeprazole) are an example of a drug class that is unstable in the gastric acid and are administered orally as an enteric-coated tablet. In addition, although some drugs are absorbed, they are metabolized by the enzymes in the gut wall or the liver prior to reaching the systemic circulation. Lidocaine is an example of a drug that is metabolized so extensively as it passes through the liver following oral absorption that effective systemic effects require parenteral administration. Extensive hepatic metabolism following oral absorption is referred to as a *First Pass Effect* (see Chapter 58 *Hepatic Clearance*). Recently a greater appreciation for the impact of drug transporters on oral bioavailability of a number of compounds has been realized. In particular, the xenobiotic transporter P-glycoprotein has been shown to significantly affect the oral bioavailability of cyclosporine and other large hydrophobic compounds. Similar to the knowledge gained by studying the CYP450 enzymes responsible for metabolism of commonly prescribed drugs, knowledge of the substrate specificity of P-glycoprotein is integral to predicting the bioavailability of drugs that are substrates for this transporter.

Bioavailability or F, refers only to the extent of absorption. The rate of drug absorption can also be in important factor in drug administration. Extended release tablets and capsules are often designed for the drug to be slowly released from the dosage form so that drug absorption is relatively constant over the entire dosing interval. As a result these types of oral dosage forms tend to produce relatively little fluctuation in the plasma drug concentrations within a dosing interval. While this may be ideal for a drug with a narrow therapeutic index, these drug products may not be useful when relatively rapid drug onset is desired. In addition the drug release characteristics are usually designed to be consistent with a specific dosing interval. If a drug product is designed to be absorbed over 12 hours, extending the doing interval to 24 hours may result in unacceptable swings in plasma concentrations.

Patients with certain gastrointestinal diseases may have a very short gastrointestinal transit time and thereby limiting the use of some extended release drug products. One example of a slowly absorbed drug with a limited bioavailability is phenytoin in the newborn. While not designed as an extended release product, phenytoin is so limited in its water solubility that several hours are required for complete absorption following oral administration. The newborn child has such a short GI transit time that when infants are changed from parenteral to equal oral doses of phenytoin, the plasma concentrations almost always decline dramatically because of a limited oral bioavailability.

Volume of Distribution (V)

Following absorption, drugs distribute throughout the body. Each drug has its own characteristics that result in an apparent volume of distribution (V) and can be expressed mathematically as:

$$\text{Volume of Distribution} = \frac{\text{Amount of Drug in the Body}}{\text{Plasma Concentration}}$$

or

$$V = \frac{\text{Amount of Drug in the Body}}{C} \quad (5)$$

where V is the volume of distribution and C is the plasma drug concentration. As can be seen from the equation above, volume of distribution is the volume required to account for the drug assuming the tissues have the same concentration as plasma. Volume of distribution is an important pharmacokinetic parameter when calculating the loading dose required to rapidly increase the plasma drug concentration to some desired concentration:

$$\text{Loading Dose} = \frac{(C)(V)}{F}$$

where C is the desired plasma concentration and F the bioavailability. In some cases, there may be drug already present and only a partial or incremental loading dose is needed to achieve the desired C_{Target}.

$$\frac{\text{Incremental}}{\text{Loading Dose}} = \frac{(C_{Target} - C_{Initial})(V)}{F} \quad (6)$$

In the above equation C_{Target} is the desired concentration following and $C_{Initial}$ is the drug concentration just prior to the incremental loading dose.

Body Composition and Volume of Distribution

Volume of distribution is most often reported as L/kg. The applicability of this L/kg value assumes that the physical characteristics of the patient are similar to the study population. Patients who are obese, emaciated, or have extensive third spacing of fluid (ascites or edema) may have an altered volume of distribution based on total body weight. Therefore some assessment of body composition is important when making initial estimates of V.

OBESE VERSUS IDEAL BODY WEIGHT—When patients are obese the most common approach is to calculate the patient's Ideal Body Weight (IBW):

$$IBW_{males} = 50\ kg + 2.3(\text{Height in inches} > 60) \quad (7)$$

$$IBW_{Females} = 45\ kg + 2.3(\text{Height in inches} > 60) \quad (8)$$

IBW in the above equations is in kg, and it is this weight that is generally assumed to represent a "non-obese" weight. When the volume of distribution is known to correspond best to ideal or non-obese weight, it is the IBW that should be used for obese patients. As a practical approach, if a patient who weighs more than their IBW, most clinicians consider the patient to be clinically obese only if the patient is greater than 120% of their IBW:

$$\text{Clinically Obese} = \left(\frac{\text{Patient's Weight}}{IBW}\right)100 > 120 \quad (9)$$

There are a few drugs that either part or all of the excess adipose weight in the clinically obese patient is used in calculating the apparent volume of distribution. Care should be taken to

carefully evaluate the patient's weight as well as the characteristics of the specific drug in question.

EXCESS THIRD SPACE FLUID (EDEMA AND ASCITES)—Some patients have extensive edema or ascites. This fluid accumulates in the interstitial space between the vasculature and the intracellular compartment or the peritoneal cavity. The degree to which a patient's vascular volume and/or intracellular volume can change is limited. Therefore, in most cases, significant changes in body water occur in the intraperitoneal and interstitial or third space. Depending on the drug's distribution characteristics the presence of third space fluid may alter how the apparent volume of distribution is calculated. In most cases, the presence of third space fluid is evaluated by changes in weight, with 1 kg of weight gain representing 1 liter of third space fluid. Alternatively, an experienced clinician can often approximate in patients with ascites or edema the number of excess third space liters present.

One method that can be used to account for any third space fluid is to calculate the contribution that one would expect for each liter (kg) of excess edema or ascites. The apparent V for each liter can be calculated by multiplying the fraction of unbound drug in plasma (fu) times the number of liters of excess third space fluid.

$$V_{\text{Excess 3rd space fluid}} = (\text{fu})(\text{Liters of Excess } 3^{\text{rd}} \text{ Space Fluid}) \quad (10)$$

The units of V are liters. The liters of excessive third space fluid gain are usually estimated by subtracting the patient's current weight from their usual weight in kilograms. Care should be taken to evaluate whether or not the weight gain is in fact *excess* third space fluid. Usually this is accomplished by determining the time course of the weight gain. Muscle mass and adipose weight gain generally takes many months, but third space weight gain can occur over weeks, days, or even a few hours. The presence of or change in the patient's edema or ascites is also a factor that should be considered when estimating excess third space fluid weight. As an example, a patient who gains 10 kg of weight in 2 days may have been initially dehydrated and simply replaced a fluid deficit rather than have gained 10 L of *excess* third space fluid. On the other hand if the patient has extensive edema before gaining the 10 kg, the amount of excess third space fluid may be much more than the most recent 10 kg weight gain would suggest.

In most cases the amount of excess fluid gained is in the range of 5 to 10 L and is seldom more than 20 L. Because the contribution of 5 to 20 L is not significant for most drugs, the weight used in calculating the volume of distribution need only consider the patient's usual weight. However, if the volume of distribution is small and plasma protein binding is low (fu approaches 1), then excess third spacing of fluid should be considered in the calculation. This would be accomplished by first calculating the patient's weight without the excess third space weight and using the *non-excess third space weight* to calculate the patient's V in the usual way. In addition, Equation 10 above can be used to determine the additional contribution of $V_{\text{Excess 3rd space fluid}}$. The sum of these two values would be the most reasonable value to use for the patient's volume of distribution.

Digoxin and aminoglycoside antibiotics are two drugs that represent the extremes. Digoxin has a fu of approximately 0.9 and a V of approximately 500 L (7 L/Kg). If a patient accumulated 10 liters of excessive third space fluid, the increase in V would only be 9 L [ie, (0.9) (Liters of Excess 3rd Space Fluid)]. This increase in V is less than 2% of the total volume of distribution and therefore not clinically significant. It is important to note that the patient's weight without the excess third space fluid should be used to calculate the volume of distribution for digoxin and most other drugs with a large volume of distribution. Aminoglycoside antibiotics also have a fu of approximately 0.9 but the usual V is approximately 15 to 20 liters. Therefore, the increase in V of 9 L associated with 10 liters of excess third space fluid would be significant and would be incorporated in the calculation of V.

Two Compartment Volume Of Distribution

While it is often useful to think of the body as a single compartment, in reality we are made of hundreds if not thousands of individual spaces into which a drug distributes. However, for most drugs the volume of distribution can be conceptualized into two individual compartments. An initial first volume (V_1) consisting of plasma and other rapidly equilibrating tissues and a second more slowly equilibrating volume (V_2) (Fig 59-3).

LOCATION OF TARGET ORGAN—The two-compartment model has two important clinical implications. First is related to the location of the target organ for clinical response (therapeutic or toxic). Some drugs have an end organ for clinical response (efficacy or toxicity) that equilibrates very rapidly with plasma. Therefore, large doses administered rapidly into the smaller first compartment will result in elevated drug concentrations and have the potential for causing drug toxicity. It is also possible to give a smaller first dose that achieves an initial therapeutic concentration and response that is quickly lost as the drug concentration declines during distribution into the larger volume. Drugs whose target organ respond as though it were located in the initial volume of distribution must be administered in such a way as to avoid the transiently elevated drug concentrations during the administration process. This is most common when drugs are administered by the intravenous route. Most drugs have a maximum recommended rate of infusion. Usually this rate is designed to allow drug distribution to take place as the drug is being infused. Occasionally it is recommended to divide a dose into portions that are administered at set intervals, again allowing for distribution to be completed before the next part of the dose is administered. For some drugs, the intravenous administration rate has to be controlled because of an agent in the injectable dosage form that has the potential for toxicity. As an example, penicillin is most commonly available as the potassium salt. While rapid injection of penicillin itself can be potentially harmful, it is the potassium that is probably the most dangerous and the reason for controlling the infusion rate of IV potassium penicillin. When drugs

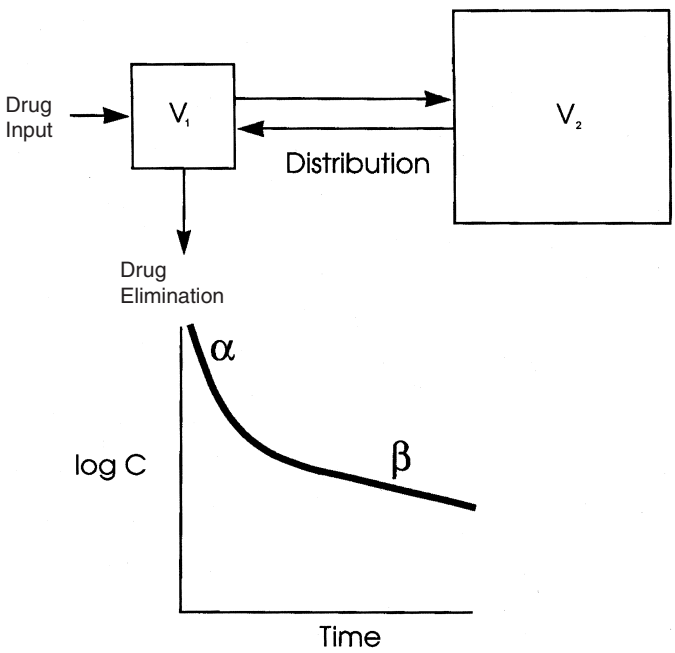


Figure 59-3. Drug first enters the body into V_1. The initial rapid decrease in drug concentration (α phase) is primarily due to drug moving into the larger more slowly equilibrating V_2. the more slowly declining drug concentrations (β phase) are primarily due to drug being eliminated from the body.

are administered orally, the absorption process is usually sufficiently slow so that elevated concentrations in the initial volume of distribution are limited. Digoxin and lithium are two exceptions, but fortunately both drugs have the end organ or response located in the tissue compartment and the transiently elevated concentrations in the initial volume of distribution do not result in an augmented clinical response.

For drugs that have the end organ for response in the deeper more slowly equilibrating volume, administration rate is not usually very critical. Digoxin is an example of a drug where the end organ for response, in this case the myocardium, is in the deep compartment. However, it is still a common clinical practice to divide large loading doses because digoxin has a significant potential for toxicity. The loading doses are divided so that the patient can be evaluated and the remaining part of the loading dose withheld if adverse events are observed.

Clearance

Clearance (CL) is a measure of the body's potential to eliminate drug. Clearance is expressed as volume/time and can be thought of as the proportionality constant between the average steady-state drug concentration (Css ave) and the rate of drug administration. At steady state the rate of drug administration must equal the rate of drug elimination.

$$\text{Rate of Administration} = \text{Rate of Elimination}$$

$$\text{Rate of Administration} = (\text{CL})\,(\text{Css ave})$$

If the Rate of Administration is expressed as F Dose/τ, where F is the bioavailability, Dose is the amount of drug administered, and τ is the interval between doses we have the following:

$$\text{F Dose}/\tau = (\text{CL})\,(\text{Css ave}) \tag{11}$$

This equation can be rearranged to calculate the desired dose necessary to achieve a desired Css ave if CL and F are known and a dosing interval τ is selected.

$$\text{Dose} = \frac{(\text{CL})(\text{Css ave})(\tau)}{\text{F}} \tag{12}$$

Alternatively, if a dose has been prescribed, the anticipated Css ave can be calculated.

$$\text{Css ave} = \frac{\text{F Dose}/\tau}{\text{CL}} \tag{13}$$

Clearance is often expressed as L/day, L/hr, or mL/min. To allow adjustment for size, clearance values are usually expressed as L/hr per kg or L/hr per m^2. There is some evidence that clearance values are best adjusted using surface area of m^2, but in clinical practice this is usually limited to patients who are substantially different from the usual 70 kg or 1.73 m^2 adult, and there is no representative patient population for a more direct comparison.

The two primary elimination pathways are elimination of unchanged drug by the renal route and hepatic metabolism.

CREATININE CLEARANCE (CLCR) AND RENAL CLEARANCE (CL_{Renal})—Renal elimination parallels renal function, and the most common measure of renal function is creatinine clearance (CLcr). The equation by Cockcroft and Gault is the most common method of estimating CLcr or renal function:

$$\text{CLcr for males (ml/min)} = \frac{(140 - \text{Age})(\text{Weight})}{(72)(\text{SCrss})} \tag{14}$$

$$\text{CLcr for females (ml/min)} = (0.85)\,\frac{(140 - \text{Age})(\text{Weight})}{(72)(\text{SCrss})} \tag{15}$$

where age is in years, weight in kg, and SCrss is the steady-state plasma creatinine in mg/dL. There are a number of assumptions inherent in the above equations. First is that the plasma creatinine is stable and not rising or falling (ie, at steady state and the patient not receiving dialysis), and the second is that the patient's muscle mass is average for his/her age, weight, and sex. In obese patients, IBW should be used in the equation to calculate CLcr. Also if a patient has extensive third spacing of fluid, that weight should not be included in the patient's weight, and for those patients who weigh less than their IBW, their actual and not IBW should be used.

Once a patient's CLcr is known or estimated, then maintenance dose adjustments can be made based on the degree of renal impairment and the fraction of the total clearance that is renal. In addition, the impact of adding to a patient's regimen a drug that can inhibit the secretion or reabsorption of a renally eliminated drug should be considered.

In some instances, it is appropriate to collect urine to directly measure creatinine clearance. The urine is collected usually over a 24-hour period, the creatinine concentration in plasma is measured, and the patient's creatinine clearance calculated by the following equation:

$$\text{CLcr (mL/min)} = \frac{(\text{U})(\text{V})}{(\text{P})} \tag{16}$$

where U is the urine concentration of creatinine in mg/dL, V is the urine volume in mL divided by the collection time in minutes. In the above equation, P is in units of mg/dL and is analogous to the value of SCrss in Equation 14 or 15 above. The value of (U)(V) is the production rate of creatinine and analogous to the 140 - Age, Weight and 72 in Equations 14 and 15. The advantage of obtaining a urine collection to measure creatinine clearance is that it is a direct measurement and does not make an assumption about creatinine production as do the methods that do not utilize a urine collection (eg, Equations 14 and 15).

There are two major disadvantages of 24-hour urine collections for creatinine clearance. The first and most obvious is that the information required to make clinical decisions may be unacceptably delayed because of the time required to collect the patient's urine. The second is that urine collections are often inaccurate. It is common for patients and even health care professionals to inadvertently discard a portion of the urine during the collection process or on occasion to collect for longer than the time listed on the collection document.

Whenever a creatinine collection is obtained, it is important to evaluate whether or not the collection appears to be adequate or complete. Although there are a number of methods that could be used, the most straightforward is to compare the CLcr from the 24-hour urine collection to the CLcr calculated from Equation 14 or 15 above. If the two values are in close agreement, it is usually a good indication that the patient is average in terms of their muscle mass, creatinine production, and the collection was properly obtained. However, if the two values are substantially different, one of the values is more likely to be the better estimate of the patient's renal function. Because both the urine collection and Equations 14 and 15 use the same value for the plasma creatinine, the difference has to be in either the accuracy of the 24-hour urine collection or the inherent assumptions in Equations 14 or 15 about the patient's muscle mass and creatinine production.

In most cases where there are significant differences, the 24-hour collection CLcr is lower than the value calculated by Equation 14 or 15. This is because the most common error is either not collecting all of the urine in the 24-hour collection period or that Equations 14 or 15 over-predicts the patient's muscle mass and creatine production. Therefore if the patient has a reasonably normal body stature and muscle mass, it is likely that Equation 14 or 15 would be the better estimate of the patient's renal function. On the other hand, if a patient is very thin and emaciated with a lower than average muscle mass, it is likely that the 24-hour collection is the better estimate of the patient's

renal function. There are other possibilities, but the examples above are the most common scenarios.

Once creatinine clearance is known, CL_{Renal} can be calculated by multiplying the patient's creatinine clearance by a "factor" or ratio of the usual drug clearance to CLcr. As an example, procainamide CL_{Renal} is about 3 times CLcr and so a patient's creatinine clearance would be multiplied by 3 to obtain an estimate of renal clearance for procainamide. Aminoglycosides have a CL_{Renal} that is approximately equal to CLcr while the factor for phenobarbital is so small that CLcr is not considered in calculating the clearance of phenobarbital.

HEPATIC CLEARANCE ($CL_{Hepatic}$)—Hepatic metabolism or $CL_{Hepatic}$ usually represents the conversion of active drug into a more polar and inactive compound by changing one or more of the functional groups on the compound. Hepatic clearance is a function of both the number and the quality of the hepatic enzymes. Clearly there are a number of factors that influence the liver's ability to metabolize drugs. Genetic composition as well as environmental factors and disease play a role in a patient's hepatic metabolic capabilities. Currently genetic profiles are not commonly available, but there are a number of drugs that are known to either inhibit or induce hepatic metabolism and we are learning more about which enzymes are responsible.

Accurate assessment of hepatic function and a patient's ability to metabolize drugs is difficult. Elevated plasma enzymes such as alanine aminotransferase (ALT), aspartate aminotransferase (AST), and alkaline phosphatase (Alk Phos) represent ongoing liver damage but are a poor reflection of function. More direct but still relatively unreliable predictors of hepatic dysfunction and decreased drug metabolism are an increased plasma bilirubin (Bili), decreased plasma albumin (Alb), and an increased prothrombin time (PT). These three biological indicators are useful in that they each represent a metabolic function of the liver. However, there are a number of factors that influence each of these laboratory tests. As an example, a patient with a low plasma albumin may have a gastroenteropathy or nephritis and have a low albumin because of a protein wasting problem and not because of decreased hepatic synthesis. Similarly, a high PT may be due to warfarin and not a reflection of a patient's hepatic metabolic capacity for drugs. Patients with severe liver failure and cirrhosis would be expected to have an elevated bilirubin and prothrombin time as well as a decreased plasma albumin. However, the presence of these abnormal laboratory test, even in the face of obvious liver disease, are not always good predictors of the extent to which a patient's ability to metabolize drugs will be decreased. Similarly congestive heart failure is known to be associated with decreased metabolism of a number of drugs. The exact mechanism of this disease-drug interaction is not known but is assumed to be secondary to either decreased hepatic blood flow or an increase in portal pressure and thereby a decrease in the ability of the hepatic enzymes to function properly.

In most clinical settings, when there is evidence of significant hepatic dysfunction the tendency is to assume that the patient's hepatic capacity is approximately half normal. Clearly this "yes" or "no" approach to estimating the presence and the extent of hepatic disease on drug metabolism leaves much to be desired. Unfortunately, with few exceptions, this is the state of the art in clinical pharmacokinetics with regard to hepatic metabolism.

OTHER CLEARANCE MECHANISMS—Renal elimination of unchanged drug and hepatic metabolism are the most common routes of elimination. However, on occasion other elimination pathways play an important role. As an example, succinylcholine is hydrolyzed and inactivated (cleared from the body in a pharmacokinetic sense) by butyrylcholinesterase enzymes located in the liver and to a significant extent in the circulating in plasma. There are other examples of in-vivo and in-vitro drug interactions that essentially act as a clearance mechanism. Aminoglycoside antibiotics with a primary amine group (eg, gentamicin) can form a covalent bond with beta-lactam antibiotics (eg, penicillin) and as a result the aminoglycoside is inactivated. This interaction is not very rapid, and significant inactivation of an aminoglycoside antibiotic only occurs when the aminoglycoside and beta-lactam antibiotic are mixed in the same IV bag or in-vivo when the two drugs are administered to a patient with end-stage renal failure. In these types of patients, the additional "clearance" can be equivalent to as much as 5 mL/min. Other potential routes of elimination are via dialysis, either hemo- or peritoneal. Some drugs are significantly removed by dialysis and care should be taken to adjust the initial dosing regimen for patients with very poor renal function and to determine if any supplemental doses would be required because of the dialysis process.

The overall approach to using pharmacokinetics and assessment of a patient's renal and hepatic function needs to be combined with the urgency of the clinical situation, the available options, and the consequences of either drug toxicity or a therapeutic failure.

Drug-Drug Interactions

With today's trend towards polypharmacy, drug-drug interactions are common and can result in therapeutic misadventures. Care should be taken to evaluate how adding or removing a drug from a patient's regimen will affect the absorption, distribution, and elimination of the other drugs the patient may be taking. Knowing the direction of the interaction (ie, increase or decrease), the time course of the onset, and the magnitude of the interaction are all important considerations when evaluating drug-drug interaction. Understanding the specifics of a drug-drug interaction helps the clinician know how and when to monitor the patient and the probable adjustments in the dosing regimen that will be necessary. As an example, both quinidine and amiodarone will approximately double a patient's steady state digoxin concentration. The doubling in the steady state digoxin concentration is due to the fact that quinidine and amiodarone both reduce by half the total clearance of digoxin; however, quinidine has a half-life of about 6 hours and accumulates to steady state within 1 day. In addition, quinidine reduces the volume of distribution digoxin, and as a result digoxin concentrations will increase rapidly within 24 hours following the initiation of quinidine therapy. Amiodarone, on the other hand, has a very long half-life and appears to have little influence on the volume of distribution for digoxin. Therefore, the increase in digoxin concentrations following the addition of amiodarone occurs over 1 to 2 weeks. Understanding these differences in how quinidine and amiodarone affect digoxin's volume of distribution vs. clearance helps to explain the difference in clinical management when either quinidine or amiodarone are added to a patient's therapy. When quinidine is initiated, a daily dose of digoxin is usually withheld to blunt the rapid rise in digoxin due to the decrease in volume of distribution, and the maintenance dose is halved to compensate for the reduction in clearance in an attempt to maintain the same digoxin concentration at steady state. In contrast, when amiodarone is added to a patient's regimen, the usual practice is to simply reduce the digoxin maintenance dose in half to compensate for the reduction in clearance.

Again knowing which parameter is affected, in which direction the parameter will be altered, and the time course as well as the expected magnitude of the change will provide the clinician with a logical approach to dealing with the drug-drug interaction.

Elimination Rate Constant (K) and Half-Life (t 1/2)

In physical chemistry, the K or rate constant is the independent parameter that controls the rate of change in a reaction. In the physiologic model used in clinical pharmacokinetics, CL and V

are the independent parameters and the relationship of CL to V that controls the rate of change of drug in the body. The fractional rate of change or K constant has the units of inverse time, usually hours^{-1} or days^{-1}) and its relationship to CL and V is represented in the following equation.

$$K = \frac{CL}{V} \quad (17)$$

or

$$CL = (K)(V) \quad (18)$$

Note that although Equation 18 appears to represent CL as a function of K and V, it is Equation 17 that represents the true dependence of K on CL and V.

Half-life is a common clinical tool and is related to K or CL and V in the following way:

$$t\ 1/2 = \frac{0.693}{K} \quad (19)$$

and

$$t\ 1/2 = \frac{0.693\ V}{CL} \quad (20)$$

The half-life of a drug is the time required for half of the drug to be eliminated from the body or the time required for a drug concentration to decline by half (Fig 59-4). This definition assumes a single compartment volume of distribution and that CL and V are constant values, independent of dose or concentration. The utility of half-life is in estimating the rate of change of drug within the body. When a drug is administered at a consistent dosing rate, the drug will accumulate to 50% of steady state after one t 1/2, 75% after 2 t 1/2 and after 3.3 t 1/2's will be at 90% of the final steady-state or plateau value. Similarly during decline, after one t 1/2 50% will be eliminated (50% remaining), after two t 1/2's 75% will be eliminated (25% remaining) and after 3.3 t 1/2's 90% will be eliminated (10% remaining). The equation that predicts drug accumulation towards steady-state assuming continuous input (see PK models section) is as follows:

$$\frac{C_t = (F)(\text{Dose}/\tau)}{CL}(1 - e^{-Kt}) \quad (21)$$

where $\frac{(F)(\text{Dose}/\tau)}{CL}$ represents the eventual Css ave, and $(1\text{-}e^{-Kt})$ represents the fraction of steady-state achieved after t hours of input at a rate of (F)(Dose/τ). Alternatively the concentration of drug remaining some time t later can be expressed by the following:

$$C_2 = (C_1)(e^{-Kt}) \quad (22)$$

where C_1 is the initial concentration, e^{-Kt} is the fraction remaining at t hours later, and C_2 is the remaining concentration. Equation 22 assumes that CL and V are constant, that the V behaves as a one-compartment model and that there is no additional drug input between C_1 and C_2. Clearly an equation can become complex when it is composed of a series of equations representing the concentration at C_1. Take for example a drug that is infused for 10 hours and then the infusion is discontinued and 5 hours later a drug concentration is to be calculated. This could be accomplished by using Equation 21 to calculate the concentration (C_t) at the end of the 10-hour infusion and then substituting that value for C_1 in Equation 22 and calculating the value of C_2 5 hours later. Alternatively Equations 21 and 22 could be combined as follows:

$$C_2 = \frac{(F)(\text{Dose}/\tau)}{CL}(1 - e^{-K\ 10\ \text{hours}})(e^{-K\ 5\ \text{hours}}) \quad (23)$$

Either approach is reasonable, but equations can become somewhat complex depending on the number of dose manipulations that are made and the time when the drug concentration is to be calculated.

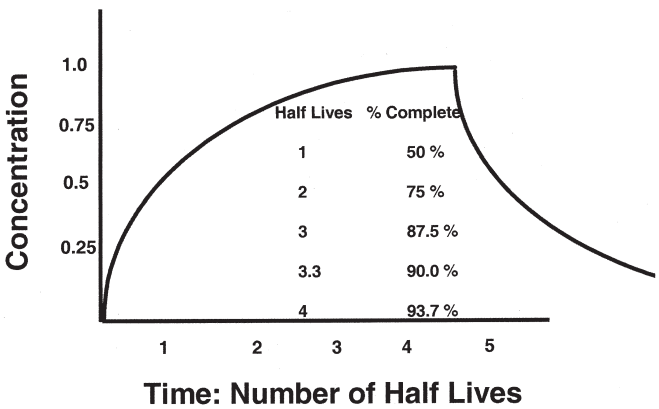


Figure 59-4. Drug Accumulation and Decline. The ascending curve represents drug concentration accumulating towards steady-state and the descending curve represents drug concentration declining after the drug is discontinued. After one t 1/2 the drug is at 50% of steady-state, two t 1/2's 75% and 3.3 t 1/2's 90-% of steady-state. In the declining phase the drug concentration declines to half of the previous concentration in each t 1/2 after 3.3 t 1/2's 90% of the drug will have been eliminated.

INDIVIDUALIZED DRUG DOSING REGIMENS

The goal of clinical PK/PD is to develop individualized drug therapy regimens to maximize the likelihood of therapeutic success. The doses utilized in clinical practice typically are derived from phase II/III clinical trials in which the safety and efficacy of an agent are evaluated. The information from clinical trials as well as preclinical work from animal and in-vitro studies provides data on the doses necessary to obtain concentrations in the "usual therapeutic range." As discussed previously, the therapeutic range defines the range of concentrations in which most patients have a therapeutic effect and a low incidence of toxicity; however, for drugs exhibiting a narrow therapeutic index, it may be necessary to more precisely define the dosage regimen for an individual based on (1) patient factors (severity of disease), (2) concentrations achievable at site of drug action (distribution, elimination), and (3) the level of sensitivity to the drug (pharmacodynamics). For example, it is widely recognized that higher doses of the aminoglycoside antibiotics are necessary in the treatment of sepsis versus a urinary tract infection. The high mortality rate from sepsis requires early aggressive therapy in order to control the infection before serious end organ damage occurs. Higher doses are also necessary where penetration into the site of the infection is reduced (ie, pneumonia or meningitis), whereas lower doses would be appropriate when renal function is compromised. These differences can be characterized and are based on knowledge of the pharmacokinetics of the drug being administered. Over the past decade, our understanding of the pharmacodynamics of drugs has improved significantly. This recognition stemmed from the observation that despite achievement of concentrations in the desired range, not all patients exhibit the same clinical response. Subsequently, a number of studies have evaluated the relationship between measures of drug exposure (Cmax, AUC), measures of disease sensitivity (eg, minimum inhibitory concentration for bacteria), and clinical outcomes (efficacy and toxicity). For the aminoglycoside antibiotics, it has been shown that the bactericidal activity and clinical improvement is correlated with the peak to MIC ratio (Fig 59-5). According to this analysis, the probability of normalizing the temperature and leukocyte count is maximized when the peak/MIC ratio exceeds 10. Therefore, higher doses would be necessary for

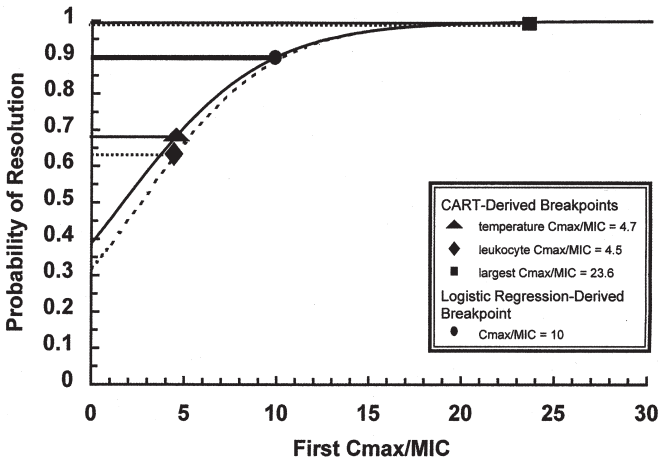


Figure 59-5. Probability of clinical improvement with aminoglycoside therapy based on the pharmacodynamic endpoint of Cmax:MIC. Resolution of temperature and leukocyte count is optimal in patients in whom the Cmax:MIC exceeds 10. (From Kashuba et al. *Antimicrob Agents Chemother* 1999; 43:623.)

treating infections involving less susceptible organisms (ie, *P. aeruginosa* vs. *E. coli*). Similarly, the probability of experiencing an adverse reaction has been linked to measures of drug exposure. The probability of a nephrotoxic event is associated with the daily AUC (>100 mg x h/L for gentamicin and tobramycin) and concomitant use of other nephrotoxic agents. In particular, in the presence of vancomycin, the amount of aminoglycoside exposure as measured by 24-hr AUC to maintain a similar risk for nephrotoxicity as aminoglycoside monotherapy is significantly reduced (Fig 59-6). Therefore, lower doses may be required when the aminoglycosides are prescribed in combination with other potentially nephrotoxic agents (ie, vancomycin). The importance of these observations is that truly individualized drug therapy cannot be established without consideration of patient factors, pharmacokinetics and pharmacodynamics, and weighing their relative importance.

Methods for Dosage Individualization

Numerous methods for dosage individualization have been described in the literature and can be grouped based on their level of precision in achieving specific concentration goals. Dosage individualization for drugs that have a relatively wide therapeutic range is typically performed using nomograms incorporating an assessment of patient demographics (ie, weight) and clinical characteristics (ie, renal function). Since renal function and weight are the principal covariates affecting the pharmacokinetics of drugs that are predominately renally cleared (ie, vancomycin), they are frequently incorporated into nomograms to reduce the variability in drug exposure between individuals within the population when compared with fixed dosing regimens (ie, vancomycin 1 gm q12h). The use of dosing nomograms such as these are most applicable for the "average" patient whose pharmacokinetics can be easily predicted based on estimates of variables such as weight and renal function. In contrast, in clinical situations where the pharmacokinetics of the drug is likely to differ significantly from the "average" patient or if the risk of therapeutic failure or toxicity is great, then a more precise measure of dosing is necessary. The aminoglycoside antibiotics are an example of a class of drugs that often meet this latter category. For many years, it has been recognized that there is a relatively narrow range of concentrations that result in efficacy and avoidance of toxicity. As a result, it is vital to determine the dose that will maximize the probability of efficacy while minimizing the risk of serious toxicity. Similar to vancomycin, weight and renal function are important variables in describing the pharmacokinetics within an individual. Therefore nomograms incorporating these factors are widely used in clinical practice to individualize the dose. However, in addition, factors such as third space fluid (edema, ascites), obesity, certain clinical conditions (ie, cystic fibrosis, burns, spinal cord injury) can alter the pharmacokinetics of the aminoglycosides. It is under these circumstances that more sophisticated PK/PD methods may be required to individualize the dosing regimen.

The use of models to determine the optimal dose using patient demographics and clinical characteristics are often referred to as *a-priori* dosing since the dosage prediction is done before direct measures of drug disposition are available (ie, measured drug concentrations). Most clinical laboratories are capable of quantifying the amount of certain drugs (drugs exhibiting a narrow therapeutic range) in biological fluids (ie, plasma), which enables a more direct assessment of the appropriateness of a dosing regimen. *A-posteriori* dosing refers to development of a revised dosing regimen based on feedback from measured drug concentrations. The drug concentration data along with the dosing history is incorporated into the PK model to determine the revised PK parameters and then are utilized to calculate a revised dosing regimen to achieve the desired concentrations.

PHARMACOKINETIC MODELS

Compartmental pharmacokinetic models, in particular the linear one-compartment open model (see *Basic PK/PD* chapter), have been extensively studied and applied to the individualization of a number of drugs used in clinical practice (aminoglycosides, procainamide, theophylline, valproic acid, vancomycin etc.). The advantage of this model is its simplicity enabling determination of dosage regimens using a handheld calculator. The disadvantage of this model is that many drugs do not exhibit instantaneous distribution. If drug levels are drawn to check the accuracy of the predictions the levels need to be appropriately timed to avoid the distribution phase. Alternatively, multicompartmental models have been employed to more accurately describe the disposition of drugs such as digoxin that exhibit a significant distribution phase. The availability of computers facilitates the relatively complex calcula-

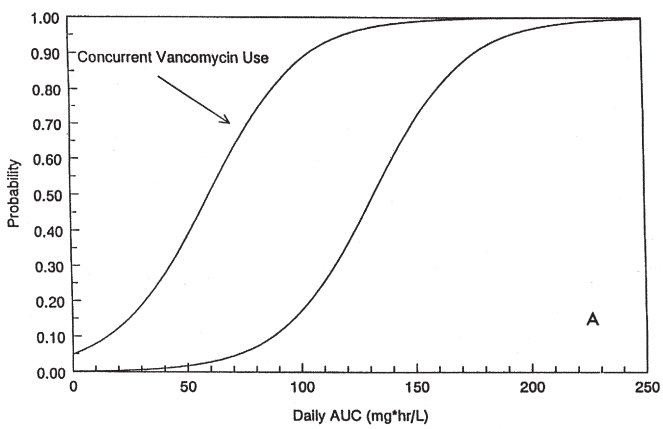


Figure 59-6. Probability of experiencing a nephrotoxic event while receiving gentamicin or tobramycin according to Daily AUC (mg x hr/L). The probability of a nephrotoxic event increases significantly when the daily AUC exceeds 100 mg x h/L. The risk is compounded by the concomitant administration of vancomycin as evidenced by a left shift in the probability curve. (From Rybak et al. *Antimicrob Agents Chemother* 1999; 43:1549.)

tions necessary to determine the revised pharmacokinetic parameters necessary to individualize the dose. The disadvantage of the multicompartment models is that typically more assay measurements are needed in order to provide good estimates of the additional PK parameters needed to describe these models. While multiple plasma concentration measurements are commonly obtained in the drug development process, typically in the clinical setting we are limited to fewer samples (eg, peak and trough). As a result, the linear one-compartment model is the most commonly employed model for predicting and revising drug dosage regimens in the clinical setting.

Steady State

In the clinical setting typically multiple doses of a medication are given to the patient, which necessitates the use of models taking into consideration potential drug accumulation. The dose and dosing interval determine the rate of drug administration [(F)(Dose/τ)] and help to establish the type of drug input process that would be most appropriate. Drugs that are administered with a dosing interval that is much shorter than the drug t 1/2 (eg, $\tau \leq$ 1/3 t 1/2) can usually be modeled as a continuous infusion as represented by Equation 13 described previously.

Continuous input

$$\text{Css ave} = \frac{\text{F Dose}/\tau}{\text{CL}} \tag{13}$$

As can be seen in Figure 59-7, when the τ is very short compared to the drug t 1/2, the difference between the peak and trough concentration is very small (<20%), and all concentrations within the dosing interval are a good representation of the Css ave. This assumption that all drug concentrations are an approximation of Css ave when $\tau \leq$ 1/3 t 1/2 is based on a one compartment model and is not valid when drug concentrations are obtained during the distribution phase.

The dosage form is also important in terms of evaluating the drug input process. For many orally administered drugs, absorption is relatively rapid, and peak concentrations will occur within 1 to 2 hours. As an example, if a drug is administered every 12 hours as a non-sustained-release product, it might be thought of as an intermittent dose and modeled with Equation 24, which is the one-compartment intermittent bolus equation that assumes instantaneous absorption and distribution. The input may be considered instantaneous even if administered orally or as a short infusion as long as the absorption/input is short relative to the elimination half-life (ie,<1/6 t1/2 ~ 10% drug loss during absorption/input). In the example, figure the drug was absorbed or infused over 1 hour and had an elimination half-life of 6 hours.

Intermittent Bolus (t_{in} <1/6 t1/2)

$$C_{ss_t} = \frac{\dfrac{(F)\text{Dose}}{V_D}}{(1 - e^{-k_{el}\tau})} e^{-k_{el}t} \tag{24}$$

where Css_t is the steady-state drug concentration at any time t after the Dose or Css max concentration.

However, if absorption is slow or a sustained-release dosage form is used, the drug input process may be more appropriately thought of as a continuous process. As an example, if a drug product is administered every 12 hours as a sustained-release dosage form with an absorption time of approximately 12 hours (time of absorption equal to τ), there would be very little change in the drug concentration within the dosing interval. The drug concentrations could therefore be modeled as a continuous infusion using Equation 13.

In contrast, the input may not be considered instantaneous when administered as a short infusion or orally if the input is considered long relative to the elimination half-life (ie, ≥1/6 t1/2). For example, if the same drug product with a 12-hour absorption time were administered with a dosing interval of 24 hours, the drug concentration time curve would be most appropriately modeled by Equation 25 as follows:

Intermittent Infusion ($t_{in} \geq$ 1/6 t1/2)

$$C_{ss_t} = \frac{\dfrac{(F)(\text{Dose}/t_{\text{in}})}{CL_T}(1 - e^{k_{el}t_{\text{in}}})}{(1 - e^{-k_{el}\tau})} e^{-kt} \tag{25}$$

where Dose/t_{in} is the rate of infusion, t_{in} is the duration of the infusion, and t is the time from the end of the infusion to when the drug concentration at steady state (Css_t) is obtained.

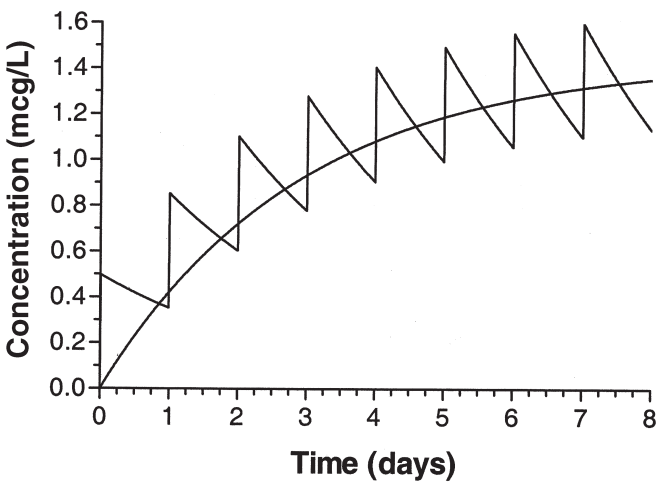


Figure 59-7. Serum concentration time profile for a drug administered as a continuous infusion. Continuous input model can also be assumed if the input time is less than one-third of a half-life.

Loading Dose

As can be seen readily from the figures (Figs 59-8 and 59-9), it may take only a few doses if $\tau \gg$ t1/2 or many doses if $\tau \ll$ t1/2 for concentrations to reach steady. Under these circumstances, it may be clinically beneficial to administer a loading dose to rapidly achieve therapeutic concentrations. As discussed previously, the incorporation of the initial concentration into the loading dose equation (Equation 6) to account for any existing drug in the body that may be present prior to administration of the loading dose should be considered.

PK/PD DOSAGE ADJUSTMENT

Drug Dosing in Renal Disease

Dosage adjustment of drugs in patients with renal impairment should be based on a knowledge of the pharmacokinetic parameters of the drug and, when indicated, on monitoring of plasma drug concentration. The aim of individualizing dosing regimens for patients with impaired elimination is to maintain plasma concentrations similar to that of patients with normal elimination and, thus, to avoid unnecessary toxicity or loss of efficacy.

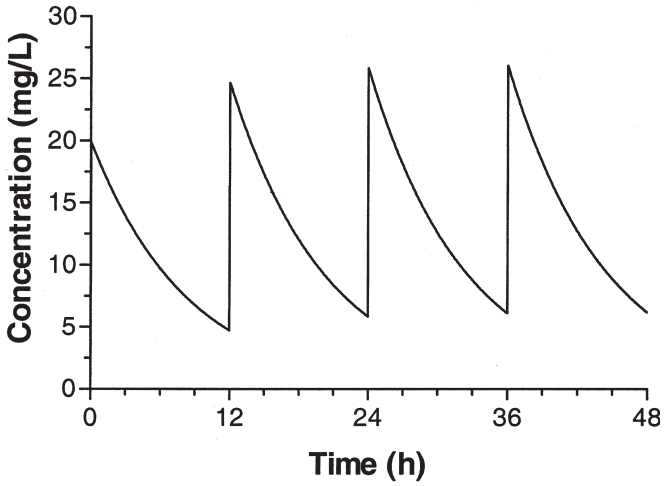


Figure 59-8. Serum concentration time profile for a drug administered as an intermittent bolus dose. Bolus input can also be assumed if the input time is less than one-sixth of a half-life.

In Chapter 58 it can be seen that $C_{ss(ave)}$ is a direct function of dose *(D)* and bioavailability (F) and an inverse function of the dosing interval (τ) and clearance. In the patient with impaired elimination or decreased clearance, $C_{ss(ave)}$ will increase until a new steady state is achieved. If clearance is impaired markedly or if the therapeutic index of the drug is small, toxicity may occur.

It is apparent from the same equation that either an appropriate decrease in dose or increase in the dosing interval will offset a decrease in elimination, and a $C_{ss(ave)}$ can be attained that is similar to that in a nonimpaired patient.

In the patient with renal impairment, individualization of drug therapy requires knowledge of the degree of impairment and its effect on drug elimination, to choose a proper dose or dosing interval to achieve a desired $C_{ss(ave)}$. As discussed above, the endogenous creatinine clearance is usually the most practical index of GFR, and it is used widely (with the limitations indicated) to determine the degree of renal impairment in a patient with renal disease.

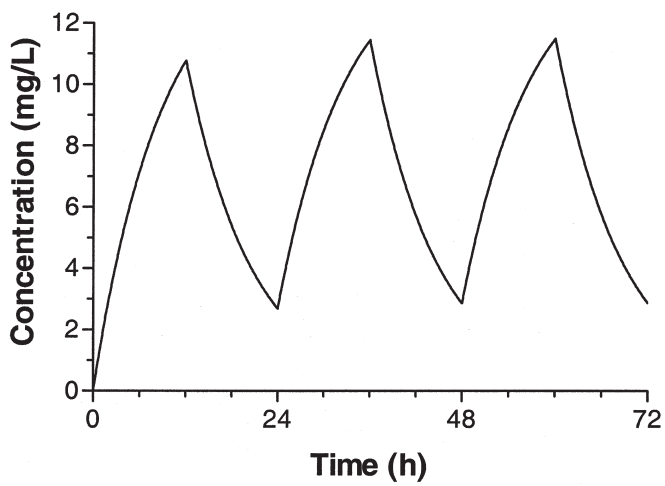


Figure 59-9. Serum concentration time profile for a drug administered as an intermittent infusion dose. The intermittent infusion model (zero order input) can also be assumed for most oral dosage forms where the input time exceeds one-sixth of a half-life.

In the literature, there are a variety of nomograms and equations available to aid in calculating dosage regimens for patients with renal impairment as discussed previously. Each is based on a set of assumptions that provide limitations to its use. Therefore, a nomogram or an equation used to determine a dose of a drug to be given to a patient with renal impairment must be used only as a guideline and, when possible, should be used along with monitoring of plasma drug concentration, when indicated, and careful clinical observation to ensure optimal therapy.

The fractional drug clearance in patients with renal insufficiency (F_{CL}) can be estimated from the relationship of the creatinine clearance in the renally impaired patient, the creatinine clearance of normal persons, and the fractional renal and nonrenal clearance of drug in patients with normal renal function according to the equation:

$$F_{CL} = \left[\frac{(CL_{cr})_I}{(CL_{cr})_N}(fe)\right] + (1 - fe) \qquad (26)$$

$$F_{CL} = \textit{(Fractional renal clearance)} + \textit{(Fractional extrarenal clearance)}$$

where F_{CL} is the fractional drug clearance in a patient with renal insufficiency, $(Cl_{cr})_N$ is the normal creatinine clearance, $(Cl_{cr})_I$ is the creatinine clearance in the patient, and *fe* (is the fraction of drug excreted unchanged in the urine in patients with normal renal function. The creatinine clearance in normal 70 kg individuals is typically in the range of 100–120 ml/min. If the patient's body size differs significantly from 70 kg then the creatinine clearance estimate should be normalized to 70 kg. Values for *fe* are readily available in the product information for most currently marketed drugs.

An example of how this PK dosage adjustment method can be applied is as follows. Insofar as the maintenance dose is concerned, the dosage regimen for the patient in renal failure can be modified by adjusting the dose, the dosage interval, or a combination of both according to the calculated dose fraction.

$$(D/\tau_{ri}) = (D/\tau)_n\,(F_{CL}) \qquad (27)$$

$$\begin{bmatrix}\text{Adjusted regimen in renal disease}\end{bmatrix} = \begin{bmatrix}\text{Regimen in normal renal function}\end{bmatrix} \times \begin{bmatrix}\text{fractional drug clearance}\end{bmatrix}$$

where (D/τ_{ri}) is the dose and dosing interval in renal insufficiency, $(D/\tau)_n$ is the usual dose and interval for normal persons, and the fractional drug clearance is the value determined from the equation as described above. An example of an adjustment in a ganciclovir dosage regimen for a patient with an impaired creatinine clearance of 35 mL/min is as follows: the usual ganciclovir dosage regimen in a patient with normal renal function is 5 mg/kg every 12 hr. The fraction of ganciclovir excreted unchanged in the urine is 0.90.

$$F_{CL} = \left[\frac{(35\ ml/\text{min})}{(120\ ml/\text{min})}(0.9)\right] + (1 - 0.9) = 0.26 + 0.1 = 0.36$$

The fractional renal and extrarenal clearance totals 0.36 indicating that the dosage regimen for this patient should be approximately one-third of the normal dose. The dose, interval, or a combination of the two can be used to determine the appropriate adjusted regimen.

$(D/\tau_{ri}) = (D/\tau)_n\,(F_{CL})$ = 5mg/kg/0.5 days (0.36) =

3.6 mg/kg/day or 1.8 mg/kg q12h

Thus, for this patient with impaired renal function, a once-a-day dose of 3.6 mg/kg or 1.8 mg/kg q12h is likely to maintain therapeutic plasma concentrations. The decision to adjust the dose or the dosage interval also should be individualized. Fluctuations in plasma concentrations of ganciclovir will be lower if

the dosage interval remains 12 rather than 24 hours. However, for some drugs there may be a therapeutic reason to achieve a higher peak plasma concentration by administering a higher dose less frequently (eg, once-daily aminoglycosides). As mentioned above, this or any other nomogram or calculation for dosage adjustment is only an approximation. Once the dosage adjustment has been made, careful clinical observation and, when indicated, monitoring of plasma concentrations is warranted. Since the loading dose depends primarily on the *V*, a change only in clearance does not typically necessitate a change in the loading dose.

Adjustments Based on Targeted Concentration and/or Pharmacodynamic Response/Surrogates

OBTAINING CLINICAL DATA—Adjustment in a patient's drug regimen requires a careful assessment of the relationship between the drug dosing regimen, laboratory data and the patient's clinical response.

DOSING REGIMEN—Knowing the dose, dosage form, frequency of administration, and duration of a drug regimen is an important piece of information. Clearly an accurate history of the patient's drug intake is one of the key elements required for pharmacokinetic interpretation. The dose of each drug administered as well as the dosage form is important. The route of administration can alter the bioavailability (F) for a number of drugs. Some drugs have a low oral bioavailability that must be considered when the route of drug administration is changed from the intravenous to the oral route.

Linking the drug input characteristics and the pharmacokinetic properties of the drug to the appropriate equation is an important step in the data gathering process.

The duration of drug therapy for a specific regimen is important in determining whether or not steady state has been obtained. As previously indicated, when a patient is receiving a dose at a constant interval, 90% of steady state will be achieved after 3.3 half-lives. While 90% of steady state is probably a reasonable approximation, many clinicians use 4 to 5 half-lives as the time required for steady state to be achieved. This is because in clinical practice the drug's t 1/2 in an individual patient may be shorter or longer than the usual value. Using 4 to 5 t 1/2 as the time required to achieve steady state helps to ensure that steady state has been achieved and decreases the chances that a false assumption about steady state will be made. (See revision of parameters.)

MONITORING EFFICACY AND TOXICITY—In the clinical setting, there are a number of factors that should be considered when evaluating a patient's response to drug therapy. Efficacy of the drug is usually focused on the disease or symptom being treated. As an example in the treatment of an infection, a reduction in a patient's fever, and a decrease in the inflammatory process would be signs of efficacy. Depending on the site of the infection, the inflammatory symptoms could range from swelling and erythema for soft tissue infections to pain or burning on urination for cystitis to mental acuity or headache for central nervous system infections. In addition, laboratory data such as a reduction in white blood cell count could also be used to monitor efficacy. For patients who are being treated for arrhythmias, suppression of the arrhythmia is often the goal and can be monitored by something as simple as taking a patient's pulse and noting that the rhythm is regular (eg, even beats with no intermittent pauses) and the rate is not excessively slow (bradycardia) or rapid (tachycardia). In other arrhythmias, electrocardiograms are used to evaluate a patient's response to therapy. Efficacy in the case of seizures is often the frequency and character of the patient's seizures. Approximately half of the patients with epilepsy are seizure-free, but the other half are only partially controlled with drug therapy.

Monitoring toxicity is equally important. Most of the drug for which pharmacokinetic calculations and plasma drug concentrations are used as an aid to determining dosing regimens have a narrow therapeutic index and or have significant toxicities. For many drugs, the order of drug toxicity is not progressive from what a clinician might consider to be "mild" to "serious." As an example, while gastrointestinal symptoms (anorexia, nausea, or vomiting) are perhaps the most commonly reported digoxin toxicities, some patients may initially present with a life-threatening ventricular arrhythmia.

The aminoglycoside antibiotics are an example of a drug that requires dose adjustment in patients with altered renal function, and plasma drug concentrations can be used to help assure that the patient is not put at additional risk for further nephrotoxicity.

Optimal Sampling Times

DRUG INPUT AND DISTRIBUTION PHASE—In almost all cases, obtaining plasma samples for drug concentrations during or shortly after the administration of a drug is not advisable. When drugs are administered by the intravenous route, there is a distribution phase that is transient and will result in either invalid or at best more complicated pharmacokinetic and clinical interpretations when employing a one-compartment model. For most drugs, the distribution phase is relatively short, and distribution is complete within 1 hour after the end of the drug infusion. Digoxin is an exception and following IV administration at least 4 hours is required for equilibrium to be attained between the plasma and more slowly equilibrating deeper compartment. Some drugs may equilibrate more rapidly and aminoglycoside antibiotics following a 30-minute infusion will distribute within 30 minutes. Clearly, sampling during the IV administration of a drug is not advisable and results in drug concentrations that are, in the clinical setting, useless.

Most drugs following oral administration are absorbed at a rate that is sufficiently slow so that the two-compartment distribution phase is not observed. Two exceptions to the limited distribution phase following oral administration are digoxin and lithium. Following oral administration, digoxin requires at least 6 hours for absorption and distribution to be complete, and lithium takes even longer. In addition, following oral administration, the onset of absorption is often delayed and/or the rate of absorption is sufficiently altered so that drug samples obtained shortly after the administration of an oral dose are difficult to interpret. For this reason, for most orally administered drugs, the preferred time to sample is at the trough. On occasion sampling at the middle of the interval is acceptable and is most common when sustained-release products are used or the trough occurs at a time that would be very inconvenient to obtain a sample.

NON-STEADY STATE—In most clinical settings, routine samples for therapeutic drug monitoring are obtained at steady state or more than 3 to 5 half-lives after starting or changing the maintenance regimen. In some cases, however, it may be advisable to obtain drug samples prior to steady state. Early sampling may allow the clinician to detect the unusual patient who is accumulating the drug rapidly and will, at steady state, have very high and potentially dangerous drug concentrations or the patient who can clear the drug unusually well and would have an unnecessarily prolonged time with low and non-therapeutic concentrations.

As a general rule drug samples obtained within 2 half-lives of starting therapy are useful only for assessing the patient at that time. If the drug concentration is unacceptably high, it might indicate that drug administration should be stopped to allow the drug concentration to decline. If the drug concentration is unacceptably low, it might indicate that an incremental loading dose should be administered to rapidly move the drug concentration into the desired concentration range. If the drug concentration is within a reasonable range, however, it does not mean that the maintenance regimen is appropriate because

drug samples obtained within the first 2 half-lives of starting or changing therapy are not useful for revising clearance and predicting steady-state concentrations.

Drug concentrations obtained after 2 half-lives but before 3.3 to 5 half-lives do contain some information about clearance and steady state but generally require more complex, non-steady state pharmacokinetic calculations and are most easily done using a computer program.

STEADY STATE—At steady state, most plasma samples for routine monitoring are obtained at specific times that allow pharmacokinetic interpretation. As previously discussed, no sample should be obtained during the drug administration/-absorption or distribution time. In addition, comparing the expected or usual drug t 1/2 and the dosing interval as well as the dosage form type (rapid vs. sustained) helps to determine the optimal time for obtaining samples.

For drugs that are administered with a dosing interval that is less than 1/3 of the drug's t 1/2 or if the dosage form is designed to release the drug over the entire dosing interval (Fig 59-7), then a single sample obtained at almost any time is acceptable so long as the absorption and distribution phase is avoided. However, for simplicity these drugs are usually recommended to be sampled at the trough or just before a dose. In some cases, sampling at the trough is not convenient, and midpoint sampling may be acceptable. Under these conditions, all of the drug concentrations within the dosing interval are assumed to be a close approximation of the Css ave concentration, making clearance the pharmacokinetic parameter of interest.

For drugs that are administered with a dosing interval that is more than 1/3 of a t 1/2 but less than one t 1/2 (Fig 59-8), it is usually recommended that a single trough concentration be obtained. Additional samples only increase the cost and do not usually increase substantially the amount of pharmacokinetic information that can be determined. Pharmacokinetic manipulations to calculate clearance require a literature estimate of volume of distribution. While the peak and trough concentrations are not a good direct approximation of Css ave, it is still clearance that is the pharmacokinetic parameter most responsible for determining both the steady state peak and trough drug concentrations when the dosing interval is less than t 1/2.

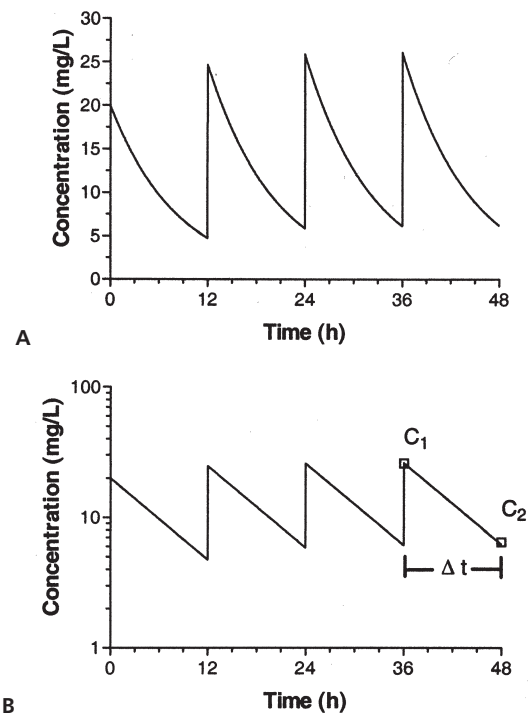


Figure 59-10. Nonlinear (a) and log-linear (b) decline in concentrations.

When the dosing interval exceeds t 1/2 and especially when the dosing interval is several t 1/2's (Fig 59-9), both volume of distribution and clearance play an important role in determining the steady-state peak and trough concentrations. Therefore, if both peak and trough concentrations are of clinical interest, two samples are required. In the clinical setting, the most common drugs following this type of dosing and plasma monitoring routine are the aminoglycoside antibiotics. In most cases, it is recommended to obtain a "peak sample" sometime after the distribution phase is complete, usually 30 minutes to 1 hour after the end of the infusion and trough concentrations within 30 minutes of the next dose. For convenience, it is common for the trough concentration to be obtained before a dose and then the peak after the dose. Although an accurate time of sampling is always appropriate, the short t 1/2 of the aminoglycoside antibiotics makes recording the time of sampling especially important.

DETERMINATION OF REVISED (A POSTERIORI) PK PARAMETERS—Once drug concentrations are obtained they need to be analyzed to determine the appropriateness of the current dosage regimen in achieving the desired goals (ie, peak, AUC). Several different methods have been described for analyzing such data and include log-linear regression, non-linear regression, and maximum *a posteriori* Bayesian analysis.

Log-Linear Regression—Log-linear regression analysis is used most widely in the clinical setting due to its simplicity requiring only a handheld calculator to determine the revised parameters. This method of analysis was first proposed for dosage individualization of aminoglycosides by Sawchuk and Zaske for use in the clinical setting. This method is based on the observation that for drugs for which the disposition can be adequately described using a one-compartment model, the concentrations decline in a log-linear relationship. As illustrated in the figure below (Fig 59-10), the concentrations decline in a nonlinear fashion when plotted on a linear scale; however, if plotted on a natural logarithm scale, the concentrations decline in a linear fashion. The importance of this observation is that the elimination rate constant can be readily determined from slope = K.

In the clinical setting typically we are able to obtain a peak and trough concentration to assess the adequacy of the dosing regimen. These concentrations can then be utilized to revise our estimates for the elimination rate constant using the aforementioned log-linear model as demonstrated below. The half-life can then be calculated from the revised *K using equation 19 as described earlier.*

Revise K and *T 1/2*:

$$\mathrm{K} = \frac{\ln \frac{C_1}{C_2}}{\Delta t} \tag{28}$$

$$\mathrm{t\ 1/2} = \frac{0.693}{\mathrm{K}} \tag{19}$$

The estimated volume of distribution can then be revised using the one compartment model substituting values for the dose, measured drug concentration, revised elimination rate constant, and time within the dosing interval that the drug concentration was obtained relative to the start of the dose. The revised clearance estimate can then be calculated from the revised K and *V*.

Revise *V* and *CL*:

$$V_d = \frac{(F)\mathrm{Dose}}{\frac{C_1}{1 - e^{-K\tau}}\, e^{-Kt}} \tag{29}$$

$$\mathrm{CL} = (\mathrm{K})\,(\mathrm{V}) \tag{18}$$

As stated earlier, the simplicity of this model accounts for its widespread use in clinical practice. One significant limitation of

this model is that it requires that the levels to be analyzed are all within the same dosing interval or are obtained at steady state with the same dosing regimen. Therefore, this method cannot be used in situations where multiple sets of drugs levels are available or when the levels are from different dosing intervals under non steady state conditions.

Nonlinear Regression—Nonlinear regression analysis is a tool available in many statistical and commercially available pharmacokinetic software programs. This method employs regression analysis on the unaltered drug concentration data (ie, no log transformation). The parameter values that are identified using this method are those that result in the minimum error between the fitted and measure concentration data. The advantage of this type of analysis is that multi-compartmental models can be utilized to fit the data if necessary. In addition, data from multiple dosing intervals obtained under steady state or nonsteady state conditions can be analyzed. One limitation to this method is that the revised parameters are determined based on the best fit to the measured drug concentration data without regard to the expected or a priori estimates of the parameters. Since many different combinations of the parameters might equally well explain the data it is possible to identify parameter values which are well described by the model but may not correlate well with values from the population particularly if the data is sparse (ie, single peak and trough concentrations).

Maximum A Posteriori Bayesian—MAP Bayesian analysis is a data analysis tool which seeks to identify the parameter values that best fit the measured data (similar to nonlinear regression) and which are most likely given the prior expectations for the values within the population. The initial parameter values utilized in the data fitting are those obtained from prior pharmacokinetic studies performed in similar types of patients. Additional iterations are performed to identify the parameter values that minimize the residual error between the fitted and measured drug concentrations as well as the difference between the revised and expected parameter values. Therefore, MAP Bayesian analysis is thought to provide the most likely estimates of the parameters. Similar to nonlinear regression, multicompartmental models can be utilized to fit the data if necessary, and data from multiple dosing intervals obtained under steady state or nonsteady state conditions can be analyzed.

CASE STUDIES

Digoxin Case History

TY is a 70-year-old, 5 foot 7 inch, 77 kg man who was admitted to the Coronary Care Unit for CHF and atrial fibrillation. He is in stable condition, but because of his rapid ventricular response it is decided that he should receive an IV loading dose of digoxin followed by an oral maintenance dose with the target digoxin concentration of 1.5 mcg/mL.

TY has lower extremity edema, and review of his previous records indicates that one month ago his weight was 69 kg.

Laboratory:

Na^+	134 mEq/L	HCT	36 %	AST	28 IU/L
Cl^-	101 mEq/L	Hgb	12.2 gm/dL	ALT	55 IU/L
K^+	4.4 mEq/L	WBC	5.4 K/mcL	TSH	3 mIU/mL
HCO_3	23 mEq/L	Plts	170 /mm^3	T. Bili	0.6 mg/dL
BUN	41 mg/dL	Albumin	3.9 gm/dL	Gluc	119 mg/dL
SCr	3.2 mg/dL				

In order to calculate the expected pharmacokinetic parameters, clearance (CL), volume of distribution (V), half-life (t 1/2), and the elimination rate constant (K) we first need to evaluate the patient's weight. At 5 feet 7 inches and 77 kg TY would not be considered to be obese. However TY's current weight of 77 kg when compared to his weight 1 month ago of 69 kg suggests that he has approximately 8 liters of third space fluid (77 − 69 = 8). The assumption that the weight gain is all excess third space fluid is based on the presence of lower extremity edema and that an 8-kg increase in either muscle mass or adipose tissue in one month is unlikely.

When calculating creatinine clearance and digoxin's CL and V, it is the non-obese, non-excess third space fluid weight that should be used. Therefore for TY, we will use his weight of 69 kg.

In order to calculate the CL and V for digoxin, we will first calculate TY's creatinine clearance using Equation 14:

$$\text{Clcr for males (ml/min)} = \frac{(140 - \text{Age})(\text{Weight})}{(72)(\text{SCrss})}$$

$$= \frac{(140 - 70\ \textit{years})(69\ \text{kg})}{(72)(3.2\ \text{mg/dL})}$$

$$= 21\ \textit{ml}/\text{min}$$

or

$$= \frac{21\ \text{ml/min}}{69\ \text{kg}}$$

$$= 0.30\ \text{mL/min/kg}$$

Using this estimate of Clcr of 0.3 mL/min/kg for T.Y. and the following equation for $V_{Digoxin}$:

$$V_{Digoxin} = [(3.12)(CL_{cr}) + 3.84](\textit{weight})$$

where $V_{Digoxin}$ is liters, CL_{cr} is mL/min/kg, weight is kg of non-obese, non-excess third space fluid weight. Substituting the appropriate values for TY, we calculate a $V_{Digoxin}$ of 330 L.

$$V_{\text{Digoxin}} = [(3.12)(0.3) + 3.84](69)$$

$$= [0.94 + 3.84](69)$$

$$= [4.78](69)$$

$$= 330\ \text{L}$$

Note that this value of 330 L for $V_{Digoxin}$ is smaller than the commonly quoted value of approximately 500 L. The lower estimate is because of TY's decreased renal function. The reason for the smaller $V_{Digoxin}$ in patients with decreased renal function is not known but is assumed to be the result of decreased tissue binding of digoxin.

We can estimate TY's $CL_{Digoxin}$ using the equation below.

$$CL_{Digoxin} = [(0.88)(CL_{cr}) + 0.33](\text{weight})$$

$CL_{Digoxin}$ is mL/min, Cl_{cr} is in mL/min/kg and weight is the non-obese, non- excess third space fluid weight. This equation is indicated for patients with congestive heart failure. Substituting the appropriate values for TY, we calculate a $CL_{Digoxin}$ of 41 mL/min.

$$CL_{Digoxin} = [(0.88)(0.3) + 0.33](69)$$

$$= [0.264 + 0.33](69)$$

$$= [0.594](69)$$

$$= 41\ \textit{mL}/\text{min}$$

While the $CL_{digoxin}$ of 41 mL/min could be used to calculate a digoxin maintenance dose, units of L/day would be more convenient, given that the dosing interval of digoxin is usually one day. The conversion to L/day is as follows:

$$CL\ in\ L/day = CL\ in\ mL/\text{min}\left(\frac{1440\ \text{min/day}}{1000\ \text{mL/L}}\right)$$

$$= 41\ \text{mL/min}\left(\frac{1440\ \text{min/day}}{1000\ \text{mL/L}}\right)$$

$$= 59\ L/day$$

Using the values for $V_{Digoxin}$ and $CL_{digoxin}$ we can now calculate the t 1/2 and K for digoxin in TY.

$$\text{t 1/2} = \frac{0.693\ \text{V}}{\text{CL}}$$

$$= \frac{0.693(330\ \text{L})}{59\ \text{L/day}}$$

$$= 3.9\ \textit{days}$$

and

$$K = \frac{CL}{V}$$

$$K = \frac{59\ L/day}{330\ L}$$

$$= 0.179\ day^{-1}$$

Now that we have the expected pharmacokinetic parameters for digoxin in TY, we can calculate a loading dose and maintenance dose to achieve and maintain a steady-state digoxin concentration of 1.5 mcg/L.

To calculate the loading dose we would use Equation 6.

$$\text{Loading Dose} = \frac{(C)(V)}{F}$$

By substituting 330 L for V, 1.5 mcg/L for C, and 1 for F, assuming the loading dose is to be administered by the intravenous route, we calculate a loading dose of approximately 500 mcg.

$$\text{Loading Dose} = \frac{(C)(V)}{F}$$

$$= \frac{(1.5\ mcg/L)(330\ L)}{1}$$

$$= 495\ mcg \text{ or } \approx 500\ mcg$$

Digoxin is one of the drugs whose end organ for response, in this case the myocardium, responds as though it were located in the deeper more slowly equilibrating tissue compartment. The loading dose is usually divided so that one-half of the total loading dose is administered followed by one fourth and then the final one-fourth. The interval between the loading dose increments is usually from 1 to 6 hours, depending on the clinical urgency. Dividing the loading dose and waiting for it to distribute into the tissue allows the clinician to evaluate the patient's clinical response (toxicity or efficacy) before subsequent portions of the loading dose are administered. If the patient developed toxicity or achieved the desired therapeutic goal before the entire loading dose had been administered the remaining part of the loading dose would be withheld.

To estimate the daily dose necessary to maintain TY's digoxin steady-state concentration at 1.5 mcg/L we would use Equation 12 below.

$$\text{Dose} = \frac{(CL)(\text{Css ave})(\tau)}{F}$$

Using the equation above to calculate the maintenance dose required is appropriate based on the t 1/2 of almost 4 days and the dosing interval of 1 day. Under these conditions, there should be relatively little fluctuation in the digoxin concentrations within the 1 day dosing interval. Substituting 59 L/day for CL, 1.5 mcg/L for Css ave, 1 day for τ, and 0.7 for F, assuming that the daily digoxin dose will be administered orally as tablets, we calculate a maintenance dose of 126.4 mcg/day.

$$\text{Dose} = \frac{(CL)(\text{Css ave})(\tau)}{F}$$

$$= \frac{(59\ L/day)(1.5\ mcg/L)(1\ day)}{0.7}$$

$$= 126.4\ mcg/day$$

Given that a daily dose of 126.4 mcg would be difficult to administer, the patient would be given 125 mcg (0.125 mg) of digoxin daily.

To monitor for efficacy the patient's heart rate and symptoms of congestive heart failure would be closely followed. In addition the patient would also be monitored for symptoms of toxicity (eg, nausea or vomiting, anorexia, visual changes, or a new cardiac arrhythmia). If digoxin concentrations are to be obtained, care should be taken to avoid the distribution phase (no sampling for digoxin levels within 4 hours of an intravenous dose or 6 hours of an oral dose). In addition because of the expected t 1/2 of approximately 4 days any sample obtained before 12 to 20 days (3 to 5 t 1/2's) should be viewed with caution as steady state may not yet have been achieved.

Aminoglycoside Case History

DS, a 65-year-old, 5 foot 5 inch, 62 kg woman, is hospitalized and recovering from total hip replacement surgery. On postoperative day 5, she develops shortness of breath and becomes febrile. A chest x-ray is performed, and she is diagnosed with a nosocomial-acquired pneumonia. She is initiated empirically on cefepime 2 gm intravenously q12h and tobramycin 440 mg intravenously over 1 hour q24h.

Pertinent Laboratory values:

BUN 14 mg/dL SCr 1.0 mg/dL WBC 16.2 K/mcL

Is the current tobramycin regimen appropriate? In order to assess the current dosing regimen it is necessary to calculate the predicted (a priori) pharmacokinetic parameters. As discussed previously, the aminoglycosides can be adequately described using a one-compartment linear model. Since aminoglycosides are almost exclusively excreted unchanged in the urine the clearance is typically approximated by creatinine clearance.

DS's ideal body weight can be estimated using equation 8:

$$\text{IBW} = 45 + 2.3\ [\text{Ht (in)} - 60]$$

$$= 45 + 2.3\ (65 - 60)$$

$$= 56.5$$

Since DS's actual weight is only 110% of her ideal body weight, it would be reasonable to assume she is not obese, and therefore we can utilize her actual body weight in estimating the clearance and volume of distribution of tobramycin.

Therefore, the tobramycin clearance can be estimated using equation 15:

$$\text{Clcr for females (ml/min)} = (0.85)\ \frac{(140 - Age)(Weight)}{(72)(SCrss)}$$

$$= (0.85)\ \frac{(140 - 65\ \text{years})(62\ \text{kg})}{(72)(1.0\ \text{mg/dL})}$$

$$= 55\ \text{ml/min}$$

This estimated creatinine clearance indicates that DS has mild renal insufficiency (normal creatinine clearance ~ 100–120 ml/min/70kg), which is most likely due to an age related decline in renal function. This estimate of renal function can be used then to estimate tobramycin clearance. The clearance is then converted from ml/min to L/hr for convenience since the dosing interval is usually 8, 12, or 24 hours.

$$CL_{\text{Tobramycin}}\ (\text{L/hr}) = \text{CLcr (ml/min)} \left(\frac{60\ \text{min/hr}}{1000\ \text{ml/L}}\right)$$

$$= 55\ \text{ml/min}\ (0.06)$$

$$= 3.3\ \text{L/hr}$$

The volume of distribution of tobramycin approximates extracellular fluid volume and therefore can be estimated based on 25% of a patient's normal weight. If the patient is significantly obese (ie, >120% of ideal body weight) or exhibits significant third spaced fluid (edema or ascites), then these need to be taken into consideration in estimating the volume of distribution as follows:

$$Vd_{\text{Tobramycin}} = 0.25\ \text{L/kg (IBW)} + 0.1\ (\text{TBW} - \text{IBW}) + 1\ (\text{kg of fluid excess})$$

where IBW is the ideal body weight (equation 8), TBW is the total nonfluid weight, and the kg of fluid excess is typically estimated from the difference between the patients current weight and admission weight.

Since DS is not obese and does not exhibit significant third space fluid, her volume of distribution can be estimated as follows:

$$V_{\text{Tobramycin}} = 0.25\ \text{L/kg [Wt (kg)]}$$

$$= 0.25\ \text{L/kg (62 kg)}$$

$$= 15.5\ \text{L}$$

The elimination rate constant and half-life can then be calculated using the following equations:

$$K(h^{-1}) = \frac{CL}{V} = \frac{3.3\ \text{L/hr}}{15.5\ \text{L}} = 0.21\ \text{h}^{-1}$$

$$T1/2(\text{h}) = \frac{0.693}{K} = \frac{0.693}{0.21\ \text{h}^{-1}} = 3.3\ \text{h}$$

The predicted steady state peak and trough tobramycin concentrations can be predicted using the intermittent bolus equation 24.

$$C_{SS_1} = \frac{\frac{(F)(Dose)}{V}}{(1 - e^{-K\tau})}\, e^{-Kt_1}$$

where Css_1 is the steady state plasma concentration at time (t_1) from the end of infusion, to the time of sampling, and τ is the dosing interval.

$$Peak = \frac{\frac{(1)(440mg)}{15.5L}}{(1 - e^{-(0.21\text{h}^{-1})(24\text{h})})}\, e^{-(0.21h^{-1})(1\text{h})} = 23.1\ \text{mg/L}$$

This peak concentration is therapeutic based on the target of > 10 times breakpoint for susceptibility; MIC = 2mcg/mL. The trough concentration can be calculated using the same equation or by decaying the peak concentration to the time of the trough assuming monoexponential decay.

$$\text{Trough} = \text{Peak}\ e^{-\text{Kt}}$$

$$\text{Trough} = 23.1mg/L(e^{-(0.21h-1)(23\text{h})}$$

$$\text{Trough} = 0.17\ \text{mg/L}$$

Trough concentrations <0.5 mcg/mL are typically below the limit of assay detection and are therefore not useful for assessing the degree of drug exposure. Therefore, midpoint or concentrations obtained greater then 2 to 3 half-lives from the peak concentration may be more useful is determining the level of drug exposure (ie, AUC_{24}).

$$AUC_{24} = \frac{Dose_{24}}{CL}$$

$$AUC_{24} = \frac{440mg}{3.3L/hr} = 133\ \text{mg} \times \text{h/L}$$

This measure of drug exposure exceeds the target of 70–100 mg x h/L and may place the patient at increased risk for toxicity. Exposure for elderly with seemingly normal renal function (normal SCr) is much greater than for younger patients given the equivalent doses due to age-related decline in renal function. While extended interval dosing may still be appropriate, the actual dose administered may need reduction to achieve similar levels of drug exposure. Since the dose and concentrations are linearly related, a revised dose and peak estimate can be determined using a proportional adjustment.

To achieve a target AUC_{24} of ~100 the dose should be:

$$\text{C}_{New} = \text{C}_{Observed}\left(\frac{Target\ AUC}{Observed\ AUC}\right)$$

or

$$Dose_{New} = Dose_{Old}\left(\frac{Target\ AUC}{Observed\ AUC}\right)$$

$$Dose_{New} = 440mg\left(\frac{100mgxh/L}{133mgxh/L}\right)$$

$$= 330\ \text{mg given intravenously q24h}$$

$$C_{New} = 23.1mg/L\left(\frac{100mgxh/L}{133mgxh/L}\right) = 17.3\ \text{mg/L}$$

CHAPTER 60

Principles of Immunology

Susie H Park, PharmD
Stan G Louie, PharmD

Our lives are filled with stresses and problems, and yet the effort to perform the cellular and biochemical processes required for survival are closely regulated. The regulation of normal physiological activity is called homeostasis, which is activated or "inducible" in times of stress and down regulated to basal levels when the stress is eliminated. One component of this complex system is a defense mechanism more commonly known as the "immune system."

The immune system is a network of defense mechanisms classified as the humoral and cellular compartment. The humoral and cellular compartment can be further subdivided into three distinctive classes know as antibodies, complement factors, and cytokines. However the humoral compartment cannot function alone, rather, it must coordinate its activities with the cellular compartment. The combination of the two compartments will orchestrate an effective defense against any foreign intrusion. The cellular compartment is often divided into regulatory and effector components. The regulatory role is primarily carried out by T-lymphocytes bearing the cell differentiating cluster-4 (CD4+), which elaborates cytokines that in turn can regulate immune function. The effector component of the immune system is made up of cytotoxic T-lymphocytes (CTLs) and natural killer cells (NKCs) where they have the capacity to kill foreign organisms (Table 60-1).

Cellular immunity is not restricted to lymphocytes. Close coordination with myeloid cells such as macrophages, neutrophils, basophils, and eosinophils is necessary in response to specific antigenic challenges. Macrophages function primarily as scavengers, seeking out antigens that have traversed the barrier defenses such as the skin and mucosal membranes. Macrophages are important in the eradication of fungal and bacterial infections, and control of tumor proliferation. Neutrophils, basophils, and eosinophils all have granules in the cytoplasm and are grouped as granulocytes. Neutrophils are important in suppressing bacterial and fungal infections, where a reduction in the number of neutrophils will predispose the individual to life-threatening infections. Eosinophils are important components in response to allergens; however, they are also important in immune response against parasitic infections. This chapter will highlight the various functions of the immune system and how the various components work together to orchestrate a defense. An overview will also be presented on the biological consequence(s) that may occur when one of the compartment(s) is not functioning sufficiently.

LYMPHOID ORGANS

The various organs that make up the immune system include the bone marrow, thymus gland, spleen, and lymph nodes. These organs are connected by a network of lymphatic vessels, filled with lymphatic fluid and cellular elements that allow immune elements to circulate from one organ to another. The lymphatic system is similar to the circulatory system; however, the lymphatic fluid is devoid of erythrocytes. The lymphatic fluid contains high levels of leukocytes, important in response to local infections and antigenic intrusion.

Circulating cells found in the blood and lymphatic fluid are all derived from a common parental source, the pluripotent stem cells (PSCs), which reside in the marrow of long bone and the pelvis. It is estimated that PSCs make up only 0.1% of all the cells found in the marrow, yet they provide a continuous supply of cells found in the circulatory system. The ability to produce seemingly unlimited number of cells is achieved through a unique capacity called *self-renewal*. In this process, the stem cell is able to divide into two daughter cells, where one cell will further undergo the maturation and differentiation process to form circulating cells. In contrast, the second daughter cell will maintain quiescent and rejoin the pool of stem cells. Self-renewal will maintain the number of parental cells and allow the stem cells to constantly replenish the various cells found in the circulation.

A more intriguing question is what regulates the type of cells being produced. It appears that the maturation and differentiation processes are under strict control of hematopoietic cytokines or growth factors. Cytokines can influence the formation and function of either myeloid or lymphoid progenitor cells, where the binding of one cytokine can down regulate other cytokine receptors, thus preferentially regulating the maturation process. The maturation of lymphoid progenitor is influenced by the presence of lymphokines and interferons, whereas the formation of myeloid progenitor cells are influenced by programmed myeloid growth factors or colony stimulating factors (CSFs) such as erythropoietin (EPO), thrombopoietin (TPO), granulocyte-macrophage CSF (GM-CSF), granulocyte-CSF (G-CSF), and macrophage-CSF (M-CSF) (Table 60-2).

The thymus is the primary lymphoid tissue that regulates differentiation and maturation of lymphocytes. In this role, immature lymphocytes (CD3+) that enter into the circulation migrate into the thymus. In the thymic environment, a number of cytokines, growth factors, and interactions with the basement membrane will initiate cellular maturation and differentiation.

In organ ablation studies, the specific role of the thymus has been delineated. In mature mice whose thymus was removed, profound cellular immunodeficiency developed. Significantly lowered number of circulating lymphocytes was a hallmark of these mice when compared to mice with intact thymus. In genetically engineered-athymic mice that were transplanted with thymus, functional T-lymphocytes were identified in the blood.

Table 60-1. Types of Lymphocytes

TYPES OF LYMPHOCYTES	SURFACE MARKER	FUNCTIONS
T-lymphocytes		
Helper T-lymphocytes	CD4	Regulate the activation of the immune cascade
• TH_1	CD4	Regulate cellular immunity
• TH_2	CD4	Regulate humoral immunity
Suppressor T-lymphocytes	CD8	Down-regulate the immune cascade
Cytotoxic T-lymphocytes	CD8	Cellular cytotoxic activity
Natural killer Cells	NK 1.1	Antibody dependent cellular cytotoxic
B-lymphocytes	CD19 CD24	Antibody production

However, mice that received shammed transplants developed viral infections and malignancies.

The spleen is the largest lymphatic organ in the body and is located just below the diaphragm stretching from the middle to the left side of the abdomen. It serves as a filter, where reticular and macrophage-like cells line the vascular sinusoids. The spleen plays an important role in host defense against microorganisms that have penetrated barrier defense. Antigens that are found in the circulation will be filtered within the confines of the spleen. While in the spleen, the antigen will encounter a rapid and intense immune response. In addition to its ability to remove antigens, the spleen is important in eliminating old circulating cells. The numbers of B-lymphocytes, the cells that produce antibodies, found in the spleen explain why it is an important antibody-producing organ.

Lymphatic fluid circulates through the lymphatic vessels and a series of bean-shaped lymphatic tissue called lymph nodes. Lymph nodes are comprised of lymphatic vessels that lead into a connective tissue network that is filled with lymphocytes and macrophages. These connective tissue complexes are produced by reticular cells, which are specialized fibroblasts. Similar to other lymphatic tissues, lymph nodes function as a biological filter, where lymphatic fluids flow through them. Antigens and microorganisms found in the lymphatic fluid will be trapped within the connective tissue lattice. The presence of foreign intrusion will activate lymphocytes and macrophages residing within the confines of the lymph nodes and will induce proliferation of the cells and activate the inflammatory process. In the event of intense immune response, there can be noticeable enlargement of the lymph node and is referred to as lymphadenopathy.

HUMORAL IMMUNITY

Antibodies

The presence of antibodies in all fluids and secretion demonstrate its role in preventing antigenic intrusion. Antibodies are glycoproteins that can neutralize any foreign antigen. Immunoglobulins or antibodies exist in two forms, cellular and soluble forms. The cell-associated antibodies are expressed on the surfaces of resting B-lymphocytes and serve as antigen receptors. In contrast, the soluble form neutralizes foreign agents and activates the immune cascade.

CLASSES OF ANTIBODIES—Antibodies can be divided into five categories that are designated as immunoglobulin A (IgA), immunoglobulin D (IgD), immunoglobulin E (IgE), immunoglobulin G (IgG), and immunoglobulin M (IgM). The difference among the various immunoglobulins is in the nature of polypeptides that makes the entire complex. Immunoglobulin typically has identical heavy chains but different light chains. This difference in immunoglobulin structure is called isotypic variation. However, immunoglobulins from different individuals may be different due to genetic variations. These changes are referred to as allotypic variation and are usually minor and involve only one or two amino acids along the constant region (Fc) along the immunoglobulin complex. Although not important in terms of immune response, they are important as markers for the study of immunogenetics and for the detection of genetic diseases.

IgA is the major secretory antibody and is found in all physiological fluids such as tears, saliva, gastrointestinal fluids, milk, and mucous. IgA neutralizes microorganisms and toxins before such pathogens can cross epithelia. IgD exists predominantly on the surface of B-lymphocytes, and is present in very low concentration (<0.1 mg/mL) in the serum. The half-life of IgD is less than 3 days. The physiological role of IgD is not well delineated, but these antibodies act like antigenic receptors able to stimulate B-lymphocyte after antigen binding. In contrast, IgE is found almost exclusively on the surface of mast cells. Upon binding to an antigen molecule, IgE can cross-link to each other on the surface of the mast cell and stimulate the release of many allergic mediators. In addition, IgE is elevated when the host encounters a parasitic infection.

IgG is the most abundant immunoglobulin found in the serum, where normal serum concentration is 15 mg/mL, accounting for 75% of total serum immunoglobulin. The half-life of IgG is approximately 3 weeks and is dependent on the presence of antigens. IgG is capable of crossing the placenta and the immature intestinal epithelium to provide immunity to the fetus and the newborn infant.

IgG is the most important antibody in the serosal immunity. IgG has a high affinity to antigen and IgG-antigen complexes can be recognized by complement factors and by Fc (fragment constant)-receptors on the surface of phagocytes. In both cases, IgG binding leads to the elimination of antigen-bearing cells. More importantly, IgG binding facilitates natural killer cell (NKC) activity that is more commonly called antibody-dependent cell mediated cytotoxicity (ADCC). In this process, NKCs with receptors for the Fc region of IgG will attach onto the antibody and elicit its biological activity by secreting cytotoxins.

The polymeric IgM is found in the pentameric and hexameric form in the blood, where the structure is formed through disulfide linkages between the immunoglobulin moieties and a polypeptide J chain. Although IgM can exist in the monomeric form, this form is found only on the surfaces of B cells. IgM is

Table 60-2. Myeloid Cytokines

FACTOR	CLASS	NO. AMINO ACID	MW (KDA)	BIOLOGICAL ACTIVITY
IL-3 Multi-CSF	I	133	14–28	Influence differentiation of immature progenitors of RBC, monocytes, granulocytes, and platelets
GM-CSF	I	127	14–35	Influence differentiation of mature progenitors of RBC, monocytes, granulocytes, and platelets
G-CSF	II	174	18.8	Terminal differentiation of neutrophils
M-CSF	II	256, 554, 438		Terminal differentiation of monocytes
EPO	II	165	30	Terminal differentiation of RBC
TPO or MDGF	II	332		Terminal differentiation of platelets

lymphocytes derived from the bursa of Fabricius, thus giving rise to the name B-lymphocytes. In humans, the ontogeny of B-lymphocytes starts in bone marrow, where stem cells form lymphocyte progenitors. These progenitor cells will differentiate and mature into plasma cells. Following antigenic challenge, circulating B-lymphocytes serve as the progenitor to plasma cells and will produce immunoglobulin type M or IgM. Subsequent antigenic exposure to the same antigen can induce the expression of immunoglobulin type G (IgG), A (IgA), or E (IgE).

Other immune cells include those that are classified as myeloid cells, which include granulocytes, monocytes, erythrocytes, and platelets. These cells have a wider variety of biological activities as compared to lymphocytes. A subpopulation of myeloid cells called granulocytes and monocytes are important in host defense. Morphologically, granulocytes have pigmented granules in the cytoplasm. Other distinctive morphologies include multi-lobed nuclei. In contrast, macrophages have an unsegmented nuclei and lack granules in the cytoplasm. Biologically, granulocytes and monocytes work in concert to eliminate foreign intrusion, particularly bacteria, parasites, and fungi.

Granulocytes are subdivided into eosinophils, neutrophils, and basophils. Eosinophils have granules filled with histamine that are released during allergic reactions. The release of histamine results in vasodilatation and pulmonary constriction that prevent more antigens from entering the body. Although eosinophils play an important role in allergic reactions, they are also important in resisting parasitic infections. Similar to eosinophils, basophils provide an inflammatory response to allergic reactions, however, the exact role of these cells is still unclear.

The most prominent granulocyte is the neutrophil, which plays a crucial role in the defense against bacterial and fungal infections. When the absolute neutrophil counts (ANC) drops below 500 cells/mm^3, an individual may become more susceptible to life-threatening infections. Most notable of these infections is gram-negative bacteremia, which is responsible for the majority of deaths associated with severe neutropenia. In the presence of foreign organisms, neutrophils will produce and secrete hydrogen peroxide (H_2O_2), which has anti-infective properties. The presence of H_2O_2 will also serve as an intracellular signal and increase transcription of stimulating cytokine production, which will initiate the immune cascade. The immune system can also be activated through the presence of nitric oxide, which has vascular dilation activity. Similar to neutrophils, macrophages are essential in the eradication of bacterial and fungal pathogens. Macrophages are antigen-presenting cells (APCs), which ingest or phagocytose and break down the antigen into recognizable fragments for immune recognition. This serves as one type of activating signal to stimulate naïve T-lymphocytes.

IMMUNE ACTIVATION

Once an antigen has penetrated the barrier defense, the immune response usually results in non-specific antibodies and complement factors binding. This coating process is also known as *opsonization*, which serves two purposes: (1) neutralize the antigen and (2) recruit cells to the affected site. After the antibody has attached onto the antigen, a conformation change within the structure of the antibody will allow cells with antibody receptors to attach onto it.

Once the antigen is recognized by antigen-presenting cells (APC), primed B-lymphocytes or mature monocytes/macrophages and then internalized via phagocytosis or endocytosis, it is degraded into smaller fragments inside an intracellular compartment, an endosome. The degradation of the antigen will make it recognizable to lymphocytes when the antigen is presented along with class II major histocompatibility complex (MHC) (Fig 60-1). Antigen fragment-MHC class II

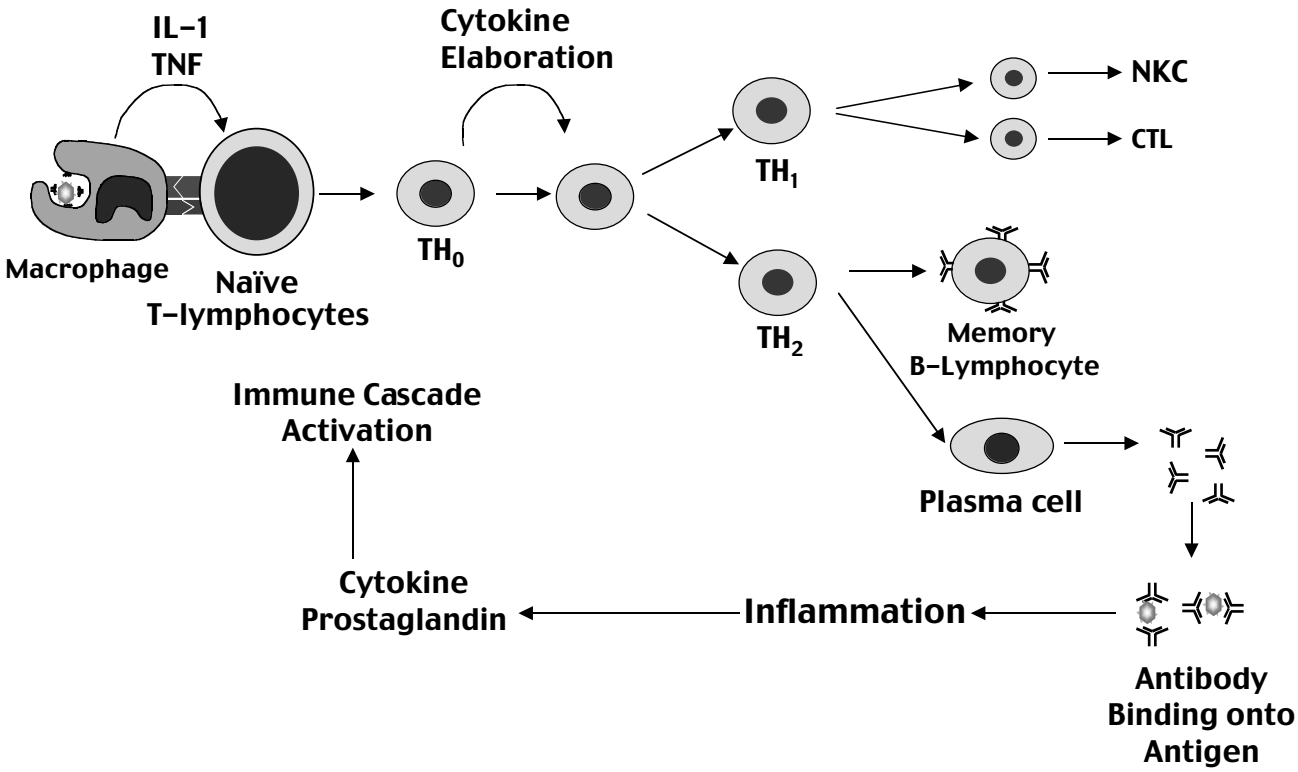


Figure 60-1. Immune activation cascade. See Color Plate 20.

the first antibody produced by the fetus and is also the first antibody to respond when presented with a new antigen challenge. In the primary response against a new antigen, the appearance of IgM in the blood precedes that of IgG. The production of IgM decreases when the production of IgG increases. Therefore, IgM in the serum amounts only about 1.5 mg/mL with a half-life in the blood of less than a week. Usually, the affinity of each antigen-binding site in IgM is lower than that in IgG because specificity has not been defined. However, the multivalent attachment of pentameric IgM provides it with very high affinity toward the surface of antigen-bearing microbials. The antigen-IgM complex can activate the complement system, but unlike IgG complexes, it cannot activate macrophage-mediated cytotoxicity.

Complement Factors

Serum has been shown to have some bacteriolytic properties; however, it was found that this antibacterial activity was lost following heat inactivation. This heat-labile activity was later called alexin. The biological activity of alexin could be restored by the addition of a second component that Paul Ehrlich called *non-immune serum*. Since alexin worked in complement with antibodies, he reasoned that alexin should be renamed as *complement factor*.

What is now known is that antibodies can recognize pathogens through antibody-antigen interaction. Antibody binding cannot only neutralize the antigen, but it can also activate the immune cascade. However, antibody binding rarely results in cytotoxic effects where the organism is killed by the addition of antibodies alone. At first, this appears incongruous with the above observation, where humoral-mediated immunity can induce cytotoxicity. It has been shown that antibodies must work in conjunction with complement factors to initiate cellular lysis of the targeted cell.

Complement factors are a group of serum glycoproteins, designated as C1 to C9. Unlike antibodies, complement factors require some assembly before it is able to carry out its functions. Complement factors will form a pore-like complex that allows intracellular contents to leak into extracellular space. The complement cascade is activated in response to ADCC. Although the complement-mediated immunity usually works in conjunction with antibodies, it can act independently of antibodies.

The complement cascade can be initiated by two distinct pathways, which are classified as the classical and alternative cascade pathways. The classical pathway is initiated when antibody binding onto the antigen causes a structural change in the Fc region of the antibody. This structural change in the Fc will activate C1. The activation of one factor will activate subsequent factors, finally forming the pore-like complex that will allow intracellular leakage to occur.

C1 consists of six C1q subunits with six binding sites at one end and is held together by the six long fibrous peptides at the other end, an arrangement that resembles six tulips held together by tightening their six stems. In addition, there are two C1r and two C1s molecules associated with the "stem" structure. The six C1q binding sites can interact with the Fc domains of antibody-antigen complexes. The six binding units in C1q can form a cross linkage between two IgG molecules. IgM molecules are the only isotype of immunoglobulins other than IgGs that are also recognized by C1q. Since IgM molecules consist of five immunoglobulin units, C1q can bind more than one Fc domain in each IgM molecule.

Upon C1q binding onto the Fc portion of IgG or IgM molecule, the C1r proenzyme is activated. Activated C1r can in turn convert C1s to an active enzymatic form, which can initiate the classic complement pathway. C4 is the first complement component that is activated by C1s. Activation of C4 will cause the glycoprotein to split and form two polypeptides, where the shorter component is referred to as C4a and the larger component is designated as C4b.

C4b is an activated enzyme capable of converting C2 to the active form, C2b. This polypeptide has enzymatic activity. When C2b is combined with C1s, C2a, and C4b, they form a complex known as C3 convertase, which converts C3 into C3a and C3b. The formation of C3b is an important step for both the alternative and the classic complement pathways. C3b is able to bind onto target cell membrane that has antibody-antigen complex together with C2a and C4b. The binding of C3b can enhance the phagocytosis of the target cell by phagocytes with C3b receptors present on the surface. The membrane bound complex of C4b, C2b, and C3b is called the C5 convertase, which produces a membrane-bound C5b fragment by cleaving off a small C5a fragment from C5 component.

During the activation process from C1 to C5, three small polypeptide fragments are generated: C3a, C4a, and C5a. While these three peptides are not directly involved in complement complex formation, they are important in the induction of inflammatory responses. C3a, C4a, and C5a are called anaphylatoxins because they are able to bind onto mast cells and basophils, thus stimulating degranulation. Mediators released by the granules can activate anaphylactic responses such as smooth muscle contraction. The anaphylactic response is inactivated by carboxypeptidase B, which removes the carboxyl terminal arginine residue of anaphylatoxins. In addition to anaphylaxis, C5a also acts as a factor of chemotaxis and an activator of neutrophils. However, the chemotactic activity of C5a is not reduced by the action of carboxypeptidase B.

Unlike factors C1 to C5, complement components following C5b in either the classical or the alternative pathway do not require modification or activation. Thus, intact C6, C7, and C8 molecules aggregate sequentially around C5b to form a membrane-associated complex, which, in turn, polymerizes C9 molecules to form a transmembrane channel. This transmembranous channel, known as the membrane attack complex (MAC), consists of an average of 15 molecules of C9 and acts as a pore to cause the leakage of electrolytes and other cytoplasmic components from the antigen-bearing cells. Eventually, the target cells are killed by this cytolytic action.

CELLS OF THE IMMUNE SYSTEM

Each cell plays an important role in maintaining homeostasis of the body. As stated above, there are various types of lymphocytes, which are divided as either B or T lymphocytes. B-lymphocytes will mature to form plasma cells that have the capacity to produce and secrete antibodies. The activity of T-lymphocytes include both regulatory and effector type activity. In contrast, myeloid cells are important in maintaining homeostasis and cell defenses.

In the early 1960s, lymphocytes isolated from lymph nodes were found to attack and destroy target cells from other animals. When these lymphocytes were introduced into tumor-bearing mice, the size of the neoplasm was reduced. In addition, these cells were able to eradicate virally infected cells, whereby giving rise to the name cytoxic T-lymphocytes (CTLs). CTLs are one of two types of effector cells that are found in humans. The second effector cell is named natural killer cells (NKCs). Unlike CTLs, this subpopulation of lymphocytes are able to attack and destroy tumor cells without prior sensitization, thus giving rise to the name *natural killer* cells. The cellular origin of NKCs has not been clearly defined. Despite this, there is little controversy regarding the role of NKC in the immune system, where the primary role of NKCs is to eliminate antibody-coated pathogen by ADCC. On the surface of NKC are receptors for the constant region of antibodies or Fcγ receptors, which allow these cells to attach onto antibody-coated or opsonized pathogenic organisms. NKCs carry out their cytotoxic function through the release of cytotoxins.

Chick embryos whose lymphoid tissue was ablated developed into chickens that were unable to produce adequate concentrations of immunoglobulins. This was attributed to a reduction of

complexes expressed on the surface of APC are easily recognized by both B- and T-lymphocytes with surface-binding antigen recognizing moieties and T-cell receptors, respectively. B- and T-lymphocytes may bind to an identical portion of the antigen fragment, but at different distinct sites. Thus, it is not unusual for most antigen molecules possessing multiple antigenic recognition sites.

Helper T lymphocytes (CD4+ cells), the cornerstone of immune activation, will only recognize antigens when they are presented on the membrane of the APC. This is to prevent producing an immune response against its own antigen. In order for the T-lymphocytes to recognize the antigen, MHC class II protein must also accompany the processed antigen. Processed antigen presentation to T-lymphocytes must also be accompanied by a co-stimulus, which includes B7 binding onto CD28. B7 is present on APC, whereas CD28 is found on the surface of CD4 cells. In the absence of this costimulation with B7/CD28, T-lymphocytes are anergic and will enter a state where it is unresponsive to antigen stimulation. In the anergic state where B7/CD28 interaction is absent, T-lymphocyte activation can utilize a secondary pathway where cytokine activation will serve as the co-stimulus.

Cytokine expression occurs as the antigen is being processed. Activation of the immune system occur when there is increased expression of interleukin-2 (IL-2) and the number of IL-2 receptors found on the surfaces of T-lymphocytes. Thus the production of IL-2 not only increases the activity of CTLs and NKC, but also of itself (autocrinic activation). This will enable autocrinic expansion, where ligand produced by the induced cells can also utilize cytokines they produce themselves. The activation of T-lymphocytes will not only increase expression of IL-2, but there is also a substantial increase in the primary inflammatory cytokines IL-1 and TNF. Similar to IL-1, IL-2 is both a paracrinic and autocrinic factor. Paracrinic factors are stimulatory cytokines that are stimulating neighboring cells. Whereas autocrinic factors are factors that are produced by the cell, which can also be utilized by the cell itself to further enhance its proliferation and/or stimulation.

There are various monocyte-derived cytokines, most notable are the primary inflammatory mediators like IL-1 and tumor necrosis factor-alpha (TNF-α). CD4+ cell activation will in turn initiate immune activation resulting in synthesis of IL-1, this will expand the number of committed progenitor stem cells. Additionally, activated CD4+ cells also produce IL-2, which will expand the population of helper-T-lymphocytes and cytotoxic T-lymphocytes, and suppress T-lymphocytes, natural killer cells, and B-lymphocytes.

There are two major types of T Helper (TH) cells that are designated as TH_1 and TH_2. TH_1 regulates cellular immunity, whereas TH_2 regulates humoral immunity. The regulation of cellular immunity is controlled by TH_1 through expression of IL-2, IL-12, and IFN-γ. These cytokines will modulate CTLs and NKCs, where IL-2 and IL-12 will increase their cytotoxic effects, respectively. IFN-γ is a cytokine synergistic with both IL-2 and IL-12, and is able to increase expression of MHC I on target cells.

In contrast, TH_2 cells regulate humoral immunity through expression of IL-4, IL-5, IL-6, and IL-10. IL-4 is able to induce the production of TH_2 cells, eosinophils, and mast cells. An increase in IgE expression is seen after IL-4 administration. In the process of activating both eosinophils and humoral immunity, IL-4 and IL-10 are also able to suppress the induction and function of TH_1 cells.

IL-1 can also stimulate the recruitment of cells to the affected site. There are two ways this can be accomplished: (1) demargination of cell immune cells adhering onto the vascular walls, and (2) stimulation of the maturation and differentiation of stem cells to increase production of circulating cells. In addition, IL-1 is also able to induce expression of CSFs and lymphokines, which regulate the differentiation, proliferation, and maturation of myeloid cells.

T-lymphocyte activation will also enhance the expression of IL-2, which was originally named T-cell growth factor (TCGF). IL-2 is a pleiotropic factor that can activate CTLs, TH lymphocytes, B-lymphocytes, and NK cells. The expression of IL-2 receptor, or IL-2R, is one indicator that the cells are activated. In order for IL-2 to exert its activity, it must first bind onto a receptor and thus activate a series of intracellular signals resulting in cellular activation.

IMMUNE DISORDERS

Inability to respond adequately to an antigen intrusion is usually considered an immunodeficiency, whereby the immune system is unable to neutralize or eliminate the pathogen. However, immune dysfunction is often the term that is used when immunodeficiency is partially impaired. Immunodeficiency occurs when one or more immune compartments are significantly affected. Immunodeficiency disorders are characterized by partial (specifically humoral or cellular) or complete impairment of immune response to an antigenic challenge. These disorders can be classified as humoral (B-cell mediated), cellular (T-cell mediated), combined immunodeficiency, phagocytic dysfunction disorders, and complement deficiencies.

The emergence of acquired immunodeficiency syndrome (AIDS) has highlighted the importance of an intact immune system. Various causes of immunodeficiency include chemical-, autoimmune-, malignancy-, and viral-induced syndromes. These disparate causes accentuate the paradigm that the various components of the immune system must work in concert to orchestrate a defense against foreign invasion.

Immunodeficiencies occur in either the humoral or cellular compartment of the immune system. Humoral immunodeficiency includes individuals who produce either inadequate quantity of antibodies or non-functional antibodies. Pre-term infants and patients with chronic lymphoblastic leukemia (CLL) are examples of individuals who may have low concentrations of antibodies or hypogammaglobulinemia. These individuals are susceptible to pyrogenic bacterial infections. Alternatively, patients with cellular immunodeficiency such as patients with cyclic neutropenia, severe combined immunodeficiency (SCID), and HIV are susceptible to non-bacterial forms of opportunistic infections. Advances in our understanding of the immune system have led to the emergence of therapies that increase patient survival.

Humoral Immune Dysfunction

Individuals who have depressed levels of either antibodies or complements are associated with increased risk of bacterial infections. The reduction of these humoral factors will impair the ability to opsonize pathogens such as *Streptococcus pneumoniae* or *Haemophilus influenzae*. Reduced levels of immunoglobulin may account for the inability to neutralize antigens and recruit cellular response. However, patients with multiple myeloma have high levels of antibodies, yet these patients are susceptible to recurrent pneumococcal infections. In this situation, the issue of antigen specific antibodies is illustrated where both antibody concentration and specificity are crucial to eliminate infections.

X-LINKED AGAMMAGLOBULINEMIA—The importance of B-lymphocytes in the immune system is highlighted in B-cell dysfunction disorders. One disorder is X-linked agammaglobulinemia (X-LA), which is an autosomal recessive genetic disorder found primarily in males. The hallmark manifestation of X-LA is low levels of IgG. This is attributed to a defect in normal lymphopoiesis. Pre-B-lymphocytes isolated from X-LA patients are unable to form mature plasma cells, thus they are unable to produce antibodies. Two possible causes of this disorder are the inability to form lymphoid progenitor cells and/or to produce inadequate cytokine signals that are required for differentiation.

Patients with X-LA are usually diagnosed relatively young, usually between 5 and 36 months of age. Clinically, these patients develop recurrent bacterial sinusitis or pulmonary infections caused by *Streptococcus* sp., *Staphylococcus* sp., *Escherichia coli*, and *Haemophilus influenzae*. Although these infections are common in this age range, X-LA patients are also susceptible to viral and protozoal infections despite having an adequate number of circulating T-cells with normal function. The inability to produce specific antibodies against a foreign organism will reduce immune response. The loss of antibody production will also reduce the cellular response and more specifically, ADCC.

SELECTIVE IGA DEFICIENCY—IgA deficiency (IgA-D) is the most common primary immunodeficiency where patients with IgA-D may not have any clinical manifestation, due to the capacity to yet compensate for IgA deficiency. Patients with primary IgA-D normally have low IgA concentration (<5 mg/dL). Inadequate IgG and IgM compensation may manifest as recurrent infections, gastrointestinal disorders, autoimmune syndromes, allergic diseases, and malignancies.

The actual pathogenesis of IgA-D is not well delineated, however, there is evidence suggesting that patients with this disorder also have a defect in HLA-A1, HLA-B8, and HLA-D expression. There is also a decrease in the number of circulating B-lymphocytes in these patients. These B-cells are able to synthesize IgA but are unable to secrete the immunoglobulins across the epithelium into extracellular space. This would suggest that IgA-D is a disorder caused by the inability to secrete the immunoglobulin.

COMMON VARIABLE IMMUNODEFICIENCY—Common variable immunodeficiency (CVID) is a primary B-lymphocyte deficiency syndrome that is characterized by low levels of serum IgG <250 mg/dL. Unlike IgA-D, low antibody levels may be seen in one or more classes. Paradoxically, low levels of IgG are not accompanied by a reduction of B-cell levels. In this disorder, circulating B-cells can be low, normal, or above normal, suggesting the cause of this disorder may be due to inadequate immunoglobulin synthesis or secretion. Other causes of CVID have been attributed to increased numbers of suppressor T-lymphocytes. This could be a consequence of the failure to adequately stimulate B-lymphocyte maturation. These findings led to the thought that CVID may be caused by viruses, such as Epstein-Barr virus (EBV). Clinical presentation of these patients support the viral-mediated etiology because these patients may develop an autoimmune defect, increases in autoantibodies, and viral-associated hypogammaglobulinemia.

CELLULAR IMMUNODEFICIENCY SYNDROMES—In contrast to humoral immunodeficiency, patients with T-cell immunodeficiencies are susceptible to viral and fungal infections. They are at risk for severe reactions to childhood disease such as varicella zoster (Chicken pox) and measles. These individuals are also at risk of developing acute infections following vaccination with live attenuated vaccines because of inadequate immune capacity to prevent subacute infection from increasing its virulence. In addition, they are more likely to develop graft versus host disease (GvHD) after transfusion contaminated with lymphocytes or allogeneic bone marrow transplantation.

DiGeorge syndrome is a T-lymphocyte immunodeficiency caused by abnormal pharyngeal pouch developmental. This normally occurs between the 6th to 10th weeks of gestation that may affect thymus, parathyroid, thyroid, heart, and certain facial features. Absence or partial absence of thymus in the newborn can inhibit T-lymphocyte development. The thymus is vital in T-lymphocyte maturation. Immature T-cells migrate from the bone marrow to the thymus. In the thymic environment, various signals and cytokines will regulate T-lymphocyte maturation and differentiation. Thus patients with DiGeorge syndrome are unable to produce mature and functional T-lymphocytes. This is evident by the lack of T-cell receptor (TCR) found in the peripheral blood lymphocytes (PBL) from patients with DiGeorge syndrome, suggesting even the most primitive T-cells are not developed due to the loss of thymic functions. Although antibody levels are oftentimes normal in these patients, they may have difficulties in producing T-cell dependent antibodies, such as specific IgG.

In contrast, Wiskott-Aldrich syndrome (WAS) is a chromosomal autosomal recessive immunodeficiency found primarily in males. WAS is caused by the inability to produce antibodies directed against polysaccharide antigens. As a consequence, abnormal granules are found in the macrophages and platelets in WAS patients. The macrophages are unable to process antigen for presentation to naive T-cells, thus leading to immunodeficiency in both B- and T-lymphocytic lineages.

At birth, WAS patients have normal lymphocyte counts but develop a decline in circulating lymphocytes with aging. More specifically, lymphocyte decline is accompanied by a drop in the number of helper T-cells along with an abnormal CD4/CD8 ratio. Furthermore, these lymphocytes do not respond to antigenic stimuli. In later stages, WAS patients can clinically present with markedly reduced T- and B-cells, low serum immunoglobulin, anergic response to exposed antigens, and recurrent infections.

Severe combined immunodeficiency disease (SCID) is another autosomal recessive disorder. Patients with SCID have profound immunodeficiency in both T- and B-cell lineages, thus have frank lymphopenia, where lymphocytes bearing CD3, CD4, and CD8 are significantly diminished or even absent. As the name describes, lymphopenia is not restricted to T-lymphocytes; circulating B-cell may also be absent. This condition will render afflicted individuals susceptible to opportunistic infections such as *Pneumocystis carinii* pneumonia (PCP). The only treatment for SCID is allogeneic transplantation for reconstituting the immune system using donor bone marrow. The identification of the specific genetic defect that causes this disorder has enabled researchers to develop methods to deliver the gene into stem cells, where the introduction of the wild type gene may ameliorate the effect of this deadly disorder.

Similar to SCID, patients with adenosine deaminase (ADA) deficiency have severe immunodeficiency. ADA is an enzyme that catalyzes the conversion of adenosine and 2'-deoxyadenosine (dAdo) to inosine and 2'-deoxyinosine, resulting in the accumulation of dAdo in body fluids and tissue. dAdo is phosphorylated to deoxy-ATP which can act as an inhibitor of DNA synthesis, resulting in cellular death. The principle site where ADA deficiency causes damage is in T-lymphocytes; however, B-cells are also affected. The molecular mechanism of ADA deficiency has been tracked to either a point mutation or deletion of the ADA gene.

The identification of a genetic mutation in patients with ADA deficiency has provided important insights into the strategies to treat this disorder. One such therapy includes bone marrow transplantation where the allogeneic stem cells have functional ADA gene. Unfortunately, allogeneic BMT has significant morbidity associated with it. Thus, other therapeutic modalities must be developed. Irradiated red blood cells (RBCs) transfusions, a rich supply of ADA, have been used to treat patients with the deficiency. Other alternatives have included bovine ADA conjugated with polyethylene glycol (PEG-ADA), which has proven to be effective in reducing levels of dAdo. PEG-ADA provides several advantages over irradiated RBCs because it eliminates the transfusion related adverse effects. In addition, PEG conjugation prolongs ADA elimination thus decreasing the frequency of administration.

CHRONIC GRANULOMATOUS DISEASE—One genetic disorder that reduces phagocytic activity is chronic granulomatous disease (CGD) syndrome, where macrophages accumulate particles in large granules in their cytoplasms. These granules are fused to each other but are unable to digest the ingested material. CGD is a rare X-linked or autosomal genetic disorder that affects phagocytic activity in the host defense. More specifically, these individuals with CGD have a defect in the NADPH oxidase system that is important in host defense against various microorganisms. The resultant effect is decreased

production of superoxide radicals, which is an important component of microbicidal mechanism. A reduction of superoxide formation can lead to recurrent serious life-threatening infections and granuloma formation. The most commonly encountered recurrent infections are catalase positive microorganisms such as *Staphylococcus aureus* and *Aspergillus* sp. Other organisms that are frequently encountered in CGD patients include *Serratia marcescens*, *Pseudomonas cepacia*, *Klebsiella* sp., *Escherichia coli* and *Nocardia* sp.

CGD is a heterogeneous disorder that is characterized by a disorder of phagocytic oxidative metabolism. CGD can occur as a result of defects in either the cytosolic or membrane component of the NADPH oxidase system. Estimates suggest that 60% of patients have a defect in the membrane oxidase system. Patients with cytosolic defects are linked to patients with autosomal recessive traits. The majority of these patients have a defect in 47 kDa cytosolic protein of the cytochrome b-558.

AUTOIMMUNE DISORDERS

The immune system normally responds specifically against foreign antigen while sparing host tissue, thus able to discriminate between self and foreign antigens. In autoimmune disorders, there are aberrations altering the ability to distinguish between foreign and self-antigens, thus permitting the immune system to attack host tissues. Occasionally, these aberrations may be initiated by infection, while at other times antigens from the infectious organisms may have structural similarities to host cellular surface markers. This can lead to immune cross reactivity, also known as immune mimicry.

There are various autoimmune disorders that are defined by the affected tissue or organ. Regardless of tissue or organ, autoimmunity is a condition where the immune system recognizes these tissues as foreign. Inflammatory bowel syndrome, systemic lupus erythematosus, diabetes mellitus, and rheumatoid arthritis are all diverse examples of autoimmune disorders (Table 60-3).

TYPE I DIABETES—Type I diabetes, or insulin-dependent diabetes mellitus, is a disease in which the selective destruction of insulin-producing β cells of the pancreatic islets of Langerhans by specific T cells results in insulin deficiency. This disease is seen almost entirely in people under the age of 30 years and peaks at age of onset between 10 and 14 years. Unlike most of the autoimmune disorders, type I diabetes occurs mostly in males.

RHEUMATOID ARTHRITIS—Rheumatoid arthritis is a complex, pathological inflammatory condition whereby autoantibodies, called rheumatoid factors such as anti-IgM, are formed against IgG. The resultant IgG:IgM immune complexes cause inflammation of the small joints of the hands and feet. This will lead to immune activation causing an increase in inflammatory cytokines such as TNF and IL-1. Individuals may experience stiffness and joint pain, accompanied by signs of articular inflammation, including swelling, warmth, erythema, as well as tenderness on palpation.

HASHIMOTO'S THYROIDITIS—Hashimoto's thyroiditis, also known as chronic thyroiditis, is an inflammatory disorder that leads to progressive destruction of the thyroid gland and symptoms of altered thyroid function. Autoantibodies to tissue-specific antigens such as thyroid peroxidase and thyroglobulin are found in very high levels in the thyroid gland for a chronic period of time. This leads to an inflammatory process that leads to fibrosis of the gland and the development of a goiter.

SYSTEMIC LUPUS ERYTHEMATOSUS—Systemic Lupus Erythematosus (SLE) is a chronic inflammatory disease that involves multiple organ systems and follows a course of alternating episodes of exacerbations and remissions. The hallmark of SLE is autoantibody production directed against double-stranded DNA. These autoantibodies are also directed against other components of the cell nucleus such as histones and ribonuclear proteins. It is the complexes of these autoantibodies and their antigen that damage tissues by activating complement. Some of the clinical features associated with SLE include the classic "butterfly" rash on the cheeks, polyarthralgia, avascular bone necrosis, myalgias, pleuritic chest pain, dyspnea, glomerulonephritis, anemia, leukopenia, and thrombocytopenia.

Table 60-3. Examples of Autoimmune Disorders

AFFECTED TISSUE OR ORGAN	AUTOIMMUNE DISORDER
Thyroid	Grave's disease or thyroiditis
Vasculature	Goodpasture's disease
Islet of Langerhans	Diabetes mellitus
Myocardial cells	Myocarditis
Platelets	Idiopathic thrombocytopenia purpura
Red blood cells	Systemic lupus erythematosus
Joints and synovium	Rheumatoid arthritis
Intestinal cells	Crohn's disease or ulcerative colitis
Skin	Dermatitis and psoriasis

HYPERSENSITIVITY AND ALLERGIC REACTIONS

Hypersensitivity reactions are immune responses to environmental antigens resulting in symptomatic reactions upon secondary exposure to the same antigen, more commonly referred to as "allergen." Hypersensitivity reactions are classified as type I to IV. Types I, II, and III are antibody-mediated reactions, whereas Type IV reaction is cell-mediated. Each type of hypersensitivity reaction, however, is unique and is summarized in Table 60-4.

Hypersensitivity Types

Type I hypersensitivity reaction is the most common category of allergic reaction and is commonly referred to as immediate or anaphylactic immune response. As the name describes, this hypersensitivity occurs after antigen (eg, pollen) binds onto IgE found on the surfaces of mast cells. Re-exposure to the same allergen will result in a cross-linking of the cell-bound IgE leading to degranulation, thus releasing its contents that include histamines and prostaglandins. Rapid release of these mediators causes profound vasodilation, increased capillary permeability, and contraction of smooth muscle. Other clinical manifestations include the development of urticaria, allergic rhinitis, angioedema, and even anaphylaxis. Systemic anaphylaxis, or anaphylactic shock, represents an extreme example of type I hypersensitivity and is considered an acute, life-threatening immunologic reaction manifesting as diffuse erythema, bronchospasm, laryngeal edema, circulatory collapse, suffocation due to tracheal swelling, hyperperistalsis, hypotension, or cardiac arrhythmias. Symptoms of anaphylaxis can develop rapidly, often reaching peak severity within 5 to 30 minutes of following allergen exposure. These clinical manifestations are primarily mediated by rapid release of eosinophil mediators such as histamine, chemotactic factor of anaphylaxis (ECF-A), and prostaglandins. Other factors include the production of slow-reacting substance of anaphylaxis (SRS-A), a group of leukotrienes produced during the anaphylactic reaction.

Type II hypersensitivity reaction is classified as cytotoxic reactions that is initiated by antibody directed against antigens found on the cell membrane of a given target cell (eg, erythrocytes, leukocytes). Antibody binding activates the complement cascade, whereby the antibody (ie, IgG or IgM) attaches to the antigen at the Fab region whereby acting as a bridge in order to complement through the Fc region. This composes a membrane attack complex which subsequently damages the cell membrane.

Table 60-4. Four Different Types of Hypersensitivity Reactions

TYPE	I	II	III	IV
Name	Immediate; Anaphylactic	Cytotoxic	Immune Complex	Delayed
Mediator	Antibody: IgE	Antibodies: IgG, IgM	Antibody: IgG	Cellular
Antigen	Atopic	Cell membrane-associated	Soluble	Tissue-associated
Target tissues	Smooth muscle	Blood	Kidneys	Varies
Target cell	Mast cells	Erythrocytes	Endothelium	Macrophages
Mediators	Histamines, leukotrienes	Complements	Complements, vasodilators	Interleukins
Mechanism	IgE antibody is induced by an allergen and binds to mast cells and basophils. When exposed to the allergen again, the allergen cross-links the bound IgE, leading to an induction of degranulation and the release of mediators (eg, histamine).	Antigens on a cell surface combine with antibody. Complement-mediated lysis then occurs.	Antigen-antibody immune complexes are deposited into tissue leading to complement is activation. Polymorphonuclear cells are attracted to the site. This causes release of lysosomal enzymes, resulting in tissue damage.	Helper T lymphocytes sensitized by an antigen release lymphokines after subsequent contact with the same antigen. Lymphokines induce inflammation and activate macrophages, which lead to the release of mediators.
Examples	Hay fever Anaphylaxis	Blood transfusion reactions	Serum sickness	Contact dermatitis
Other Characteristics	The most common form of allergic reaction			

Immune complex hypersensitivity, or Type III hypersensitivity reaction, is caused by the formation of soluble antibody-antigen complexes that aggregate in blood or tissue. These complexes adhere to various sites such as the endothelium of blood vessels subsequently leading to tissue damage. Conditions associated with immune complex include serum sickness and the Arthus reaction. Serum sickness occurs when foreign serum or serum proteins like antilymphocyte immunoglobulin derived from animals such as rabbit, goat, and horse enters the host. The recipient may then develop chills, fever, arthralgias, and nephritis. These symptoms subside as the immune system removes these agents which it recognizes as antigens. Arthus reaction is a cutaneous reaction following subcutaneously or intradermally administration leading to immune response with IgG disposition to the affected site. This leads to complement activation and phagocytic cells producing a local inflammatory response. Deposits of antigen, antibody and complement form on vessel walls, leading to polymorphonuclear cell infiltration and aggregation of platelets. Ultimately, this can lead to vessel occlusion and tissue necrosis.

Unlike the other reactions previously described, Type IV hypersensitivity reaction is a cell-mediated reaction, in particular by T-lymphocytes. Type IV hypersensitivity reaction is commonly referred to as delayed-type hypersensitivity since an immune response may not occur until hours or even days after initial contact with the triggering agent. The reaction commonly lasts several days. A classic example of this particular type of hypersensitivity reaction is allergic contact dermatitis. Antigen is processed by antigen-presenting cells and presented on MHC class II molecules, which interacts with antigen-specific T helper cells which recognized it, thus stimulates the production of IL-1 and up-regulates T lymphocyte synthesis of IL-2 and IFN-γ. These induced cytokines act on vascular endothelium and recruit the infiltration of inflammatory cells, particularly macrophages. This causes fluid and protein accumulation and consequently, local tissue destruction and lesions ensue. Acute lesions are characterized by erythema, pruritus, and vesicle formation.

Allergic Drug Reactions

Drugs are commonly implicated in causing hypersensitivity reactions (Table 60-5). Antibody-mediated hypersensitivity reactions (ie, anaphylactic, cytotoxic, serum sickness) are involved in drug allergy. Either the drug molecule itself, or its metabolite, can elicit the allergic response upon re-exposure to the identical drug. Some medications (eg, aspirin) directly stimulate mast cells. Low-molecular-weight molecules (eg, penicillin, phenytoin) become antigenic via haptenation, a chemical process by which the drug molecule reacts with host proteins in order to become immunogenic and stimulate an antibody response.

Prevention of anaphylactic reactions due to drug hypersensitivity is vital. It is important to take a complete history of a patient's past medication use and note any reactions to medications that may have induced an allergic response. Prudent clinical knowledge, and its application to medications that commonly cause allergic reactions and those agents that cross-react with them, is essential. Diagnosis of drug hypersensitivity can be accomplished via three main methods (Table 60-6): (1) in vivo skin testing for immediate reaction to a suspected agent, (2) in vitro analysis of drug-specific IgE from a person's affected blood, (3) oral challenge testing. Treatment of anaphylaxis consists of maintaining airway, breathing, and circulation control by using agents such as epinephrine and implementing supportive care.

Table 60-5. Agents Associated with Causing Allergic Reactions

AGENT/DRUG CLASS	EXAMPLES
Antibiotics	Penicillin
	Sulfonamides
	Isoniazid
Anesthetics	Propofol
Antiarrhythmics	Quinidine
	Procainamide
Antihypertensives	Hydralazine
	Methyldopa
Antipsychotics	Phenothiazine
Anti-Inflammatory Agents	Acetyl-salicylic acid
	Indomethacin
	Ibuprofen
	Naproxen
	Celecoxib
Proteins/Peptides	Insulin
Antibodies	Antisera
	Monoclonal
	Immunoglobulins
Muscle relaxants	Chlorzoxazone
	Metaxalone
Other	Monosodium glutamate

Table 60-6. Methods for Testing Drug Allergy

IMMUNOLOGIC REACTION TYPE	IN VIVO	IN VITRO
I	Immediate skin prick; intradermal	RAST ELISA
II	(none)	Coombs
III	Intradermal Arthus test	RAST ELISA
IV	Patch test	Lymphocyte proliferation

RAST = radioallergosorbent assay.
ELISA = enzyme-immunosorbent assay.

NEUROIMMUNOLOGY

The field of psychoneuroimmunology (PNI) explores the complex relationship between the nervous and immune systems. Neurology and immunology are converging as the role of neuroimmune interactions between neurotransmitters and cytokines has intertwined in health and diseases. Examples include depression, schizophrenia, anxiety, Alzheimer's disease, autoimmune disorders, chronic fatigue syndrome, stress, and sickness behavior. This section will address some of these disorders in the context of neuroimmunological dysregulation in relation to changes in behavior.

The common factors that link the two areas of study in explaining the pathophysiology of these disorders are cytokines. These pleiotropic proteins are the chemical messengers between cells that can function as both immunomodulators as well as neuromodulators. They mediate brain function as well as regulate the immune system. In addition, specific cytokines can induce the expression of neurochemical, neuroimmune, and neuroendocrine elements. Neurotropic cytokines can be secreted by cells found in the brain; these include astrocytes and microglia (immunocompetent cells within the brain). Along with secreting these cytokines, neuronal cells are also found to express receptors for the cytokines, suggesting that they are responsive to them. It was originally thought that large molecules such as cytokines could not cross the blood–brain barrier (BBB), an anatomical and functional separation between brain parenchyma and peripheral tissues that consists of vascular endothelium, basement membrane, neuroglial membrane, and glial perivascular feet. Elaborated cytokines, regardless of the source, can exert their effects on the brain via indirect and direct routes. Peripheral tissues are innervated by the peripheral and autonomic nervous systems and can send direct signals to the brain via peripheral nerves. Brain vasculature can send signals through secondary messengers such as nitric oxide (NO) or prostanoids (any group of complex fatty acids derived from arachidonic acid, including prostaglandins and the thromboxanes). These types of secondary signals are elaborated in response to cytokine activation. These secondary messengers mediate the effects of the immune molecules on brain function. Finally, cytokines can directly act on the brain by crossing the BBB or after entering an area of the brain that lacks a BBB. Cytokines released by activated immune cells influence activation of the hypothalamic-pituitary-adrenal (HPA) axis and are, in turn, influenced by glucocorticoid secretion (Table 60-7). More-

Table 60-7. Biological, Behavioral, and Psychiatric Effects of Cytokines

CYTOKINE	BIOLOGICAL ACTIVITY	PHYSIOLOGICAL EFFECTS	PSYCHIATRIC EFFECT	NEURO-TRANSMITTER EFFECTS	SECRETION SUPPRESSED BY
Proinflammatory					
TNF-α	Cytotoxic Activates T cells Pyrogenic Antitumor Septic Shock	Stimulates activity of the HPA axis	Somnolence Anorexia Cognition	Increases catecholamines	Glucocorticoids
IL-1	Activates T, B, and endothelial cells Induces acute phase proteins Pyrogenic Hematopoiesis	Stimulates activity of the HPA axis Modulates many central monoamine activity	Somnolence Confusion Delusions Sickness behavior Stress	Serotonin Dopamine Norepinephrine	
IL-6	Activates T cells Produces Immuno-globulin-G Induces acute phase proteins Pyrogenic Hematopoiesis	Stimulates activity of the HPA axis Differentiates and promotes growth of neuronal cells Increases serotonin and mesocortical dopamine activity in the hippocampus and prefrontal cortex	Somnolence Depression Psychosis Stress	Serotonin Dopamine Norepinephrine	Glucocorticoids
Inflammatory					
IL-2 (T cell growth factor)	Activates T, B, and natural killer cells Antitumor	Increases hypothalamic and hippocampal norepinephrine utilization Increases dopamine turnover in the prefrontal cortex	Depression Psychosis Confusion Delirium Memory Cognition	Dopamine Norepinephrine Acetylcholine	
IFN-γ	Activates macrophages Enhances expression of MHC Induces acute phase proteins Pyrogenic Antitumor	Neuromodulation	Fatigue Depression Suicidal ideation Psychosis Cognitive impairment	Serotonin	

over, central nervous system functioning during an immune response is modulated not only by cytokines from the periphery, but also by cytokines that are synthesized in the brain. Genes that encode for many of the cytokines and cytokine receptors are often constitutively expressed in the brain in response to immune system molecules or the cytokine itself. Within the central nervous system, cytokines can then induce changes in brain monoaminergic and cholinergic neural pathways. Cytokines modulate centrally mediated responses such as neurologic and neuroendocrine changes including the activation of the HPA axis. Cytokine receptors are localized with high densities in the hippocampus and hypothalamus of the brain.

Psychiatric Disturbances

Exogenous cytokines can influence certain behaviors such as sleep and eating, as well as modify mood states. Cytokines used therapeutically are associated with a number of adverse drug reactions including depression, mania, anxiety, irritability, decreased concentration, confusion, psychosis, and suicidal ideation. Interferons, in particular, have been implicated in causing many of these psychiatric disturbances in a dose-dependent fashion. In addition, interleukins and tumor necrosis factor-α have also been associated with causing psychological adverse effects. Since the severity of these effects appears to be dose-dependent and are reversible when the cytokine is discontinued, this would represent a causal relationship. The mechanism by which cytokines produce neuropsychiatric effects, in part, is caused by their ability to alter levels of neurotransmitters in certain areas of the brain. It is the relationship between cytokines and neurotransmitters that may explain certain behavioral disturbances.

Alzheimer's Disease

Inflammatory processes have been implicated in Alzheimer's dementia (AD) since epidemiology studies have demonstrated a lower incidence of AD in those patients using anti-inflammatory agents. Other correlative evidence includes elevated levels of inflammatory mediators in the brains of AD patients. Data suggests that inflammatory processes may initiate or enhance the pathological process that leads to the development of cerebral amyloid that may eventually lead to neuronal death. Acute phase proteins, such as α-1-antichymotrypsin, are elevated in the cerebrospinal fluid (CSF) of AD patients and appear to become integrated into the amyloid deposits that are characteristic of the pathophysiology of AD. Moreover, there are elevated levels of IL-1 and IL-6 in the serum of AD patients and IL-6 has been observed in plaques. Moreover, both of these cytokines induce the synthesis of β-amyloid precursor protein by human astrocytoma cells.

Schizophrenia

Immune abnormalities, found in some forms of schizophrenia, may reflect immunoregulation dysfunction in either the etiology or pathogenesis of this psychiatric disorder. The idea that the immune process is involved in a psychotic disorder such as schizophrenia correlates with the fact that psychosis is associated with the autoimmune disease, systemic lupus erythematosus. Schizophrenia has a chronic but episodic course similar to that seen in many autoimmune disorders. Earlier investigations have reported an elevation of serum immunoglobulin levels, abnormally large lymphocytes, and antibody to the brain in schizophrenic patients. More current research has focused on investigating specific cytokine abnormalities. Findings of altered interleukin regulation have been regarded as confirmation that schizophrenia has an autoimmune etiology, at least in part. Conclusive findings illustrate a decrease in IL-2 production in untreated schizophrenics as well as a decreased production of IL-2 correlating with clinical variables such as more negative symptoms of schizophrenia (eg, flat and inappropriate affect, cognitive deficit, alogia) and a younger age of onset of the illness. Notably, this decrease in IL-2 production is found especially in paranoid schizophrenics. This suggests that low IL-2 production occurs at an early stage in the course of schizophrenia and that IL-2 production may serve as a marker for a subtype of illness of severity of schizophrenia. Low IL-2 production may be the result of the inability of T lymphocytes to produce more IL-2 or the decreased number of T-cells that secret IL-2, as well as to the intrinsic disorder of T-cells. However, there have also been investigations that suggest an increase in IL-2 production. Furthermore, there are high concentrations of IL-2 receptors in the hippocampus and striatum; therefore it is proposed that IL-2 serves as a neuromodulator possibly by increasing dopaminergic neurotransmission. The actual effects of this postulate are unknown at this time.

In addition to changes in IL-2 production, the observation has been made that there are elevated serum levels of IL-6 in schizophrenia. There is also an association with serum IL-6 and the state of schizophrenia in which acutely ill patients appear to have higher levels than patients in remission. In order to explain decreases in IL-2 with concomitant increases in IL-6 in schizophrenic patients, it is thought that there is an imbalance with the TH_1:TH_2 with a shift to the TH_2 system. These findings that represent evidence of an autoimmune pathogenesis are gaining more acceptance. More investigations, however, are needed to ascertain the correlation between IL-2 and IL-6 changes and the severity of the symptoms of schizophrenia.

Major Depression

Evidence that immunological disturbance of acute phase proteins (ie, significantly increased haptoglobulin) and cellular immune response in patients with depression comes from the observation that there is a significant decrease in lymphocyte proliferation in response to a mitogen, in severely and moderately depressed patients. Acute phase proteins are mediated by the pro-inflammatory cytokines, mainly IL-1, IL-6, and TNF. It has been reported that serum and plasma concentrations of immunoglobulins (IgA, IgM) and C3 and C4 complement are also changed in depressed patients. Furthermore, there is much clinical and experimental data to support the relationship between cytokines and depressive symptoms, such as depressed mood, decreased appetite, anhedonia, psychomotor retardation, changes in sleep patterns, fatigue, and cognitive deficits. Depression increases the production of proinflammatory cytokines from activated macrophages in the periphery and brain, including IL-1, TNF and IL-6. The consequence of the hypersecretion of these cytokines results in the malfunctioning of noradrenergic and serotonergic neurotransmission in the brain. There is evidence that IL-1 can activate the serotonin transporter, whereby increasing reuptake of serotonin into the presynaptic neuron and decreasing the amount of serotonin in the synaptic cleft. Likewise, IL-2 alters the functional activity of the central noradrenergic system. Serotonin and norepinephrine are two of the primary neurotransmitter deficits found in depressed patients. Other findings include high CSF concentrations of IL-Iβ and lower IL-6 in depressed patients. With respect to the effects that antidepressant medications have on cytokines, selective serotonin reuptake inhibitor (SSRI) administration has been associated with an increase in IL-10, a decrease in the synthesis of IFN-γ, and a decrease in the release of IL-6. Moreover, the tricyclic antidepressant (TCA), desipramine, impairs the secretion of both IL-1 and TNF; paroxetine, an SSRI, and venlafaxine, a serotonin and norepinephrine reuptake inhibitor, do not. Whatever the alteration in cytokine production and levels, HPA axis changes induced by these cytokines are found in depressed patients. Research on cytokine regulation in depression is rather

controversial. Despite the amount of research currently done in an attempt to describe the relationship between the immune-endocrine-neurotransmitter systems, further investigation is required to measure these changes (in serum, plasma, CSF) and to correlate them to the severity of depressive illness and treatment response.

Effects and Responses to Stress

Stress in humans influences cytokine production and function by activating the HPA axis and increasing circulating glucocorticoids. It is generally stated that stress is "immunosuppressive." Specifically, induced stress and emotional distress leads to decreased IFN-γ and IL-6 and increased IL-2, as well as other effects on additional cytokines. Studies have concluded that the explanation for the changes in cytokine secretion is that stress induces an increase in the ratio TH_1/TH_2. An excessive HPA response to inflammation can mimic a condition of stress or hypercortisolemia, whereby increasing susceptibility to viral and bacterial infections

IMMUNOTHERAPEUTICS

Cancer

Immunotherapy is now an established modality in the treatment of several types of cancers, including malignant melanoma, renal cell carcinoma, multiple myeloma, and carcinoid tumor. Immunologically based anticancer therapy is based on two different types of strategies: (1) immune activation leading to specific tumor cytotoxicity, and (2) antibody therapy on tumor specific antigens.

Immunostimulants to enhance immune response against tumors have used cytokines that regulate immune response. Cytokines such as interferons (IFNs) and interleukins (ILs) are commonly used in this scenario. Although myeloid cytokines such as colony-stimulating factors (CSFs) have been investigated, their antitumor activity appears to be limited. The rationale to this disparity in biological activity is due in part to the role of cytotoxic lymphocytes (CTLs) and natural killer cells in tumor clearance. These effector cells play a major role in tumor suppression.

IFN is a family of cytokines that can induce cellular production of antiviral agents and thus block viral replication. Since a close relationship with viral infections and oncogenesis exists, it stands to reason that IFN, which can inhibit viral replication, may be able to inhibit cancer proliferation. Alpha IFN, or IFN-α, is the most frequently used biological agent used in cancer. It is approved for a broad range of cancers such as hairy cell leukemia, chronic myeloid leukemia, and AIDS-related Kaposi's sarcoma. The mechanism by which IFN-α exerts antitumor activity have included suppression of oncogenic viruses, inhibition of oncogenes, and stimulation of cytotoxic T-lymphocytes. In addition, IFN-α appears to have anti-neovascularization activity, thus it can reduce the formation of new blood vessels critical in tumor progression.

IFN-α is a key component of chronic myeloid leukemia (CML), a slowly progressing blood disorder with a number of clinical phases. Patients typically present in the chronic phase and later they develop transformation into the accelerated and blastic phase of the disease where the disease becomes resistant to conventional chemotherapy. As the disease progresses, production and function of white blood cells and platelets diminish, thus infections and spontaneous bleeding and bruising may ensue. Approximately 10% to 20% of patients who use IFN-α, alone or in combination, have complete cytogenetic response with no evidence of bcr-abl translocation (the etiological event in CML), suggesting response to therapy. These patients often times are disease-free beyond 10 years, however, maintenance of therapy with interferon is required to maintain this clinical status.

IFN-α is also used frequently in AIDS-related Kaposi's sarcoma (KS). This is the most prevalent form of KS with increasing incidence as patients become immunodeficient. KS usually presents with cutaneous lesion(s) but may present as lesions lining the mucocutaneous tracts such as the oropharynx, lungs and gastrointestinal system. Cytotoxic treatment is often given concomitantly with IFN in order to stimulate the immune system and make treatment more effective.

Recombinant human IL-2 (rhIL-2) is another lymphokine with potent antitumor activity. IL-2 is a potent stimulator of CTLs and NKCs, both responsible for immune response to the presence of cancers. Although it is approved for renal cell carcinoma and malignant melanoma, IL-2 has been used to augment a number of standard cytotoxic chemotherapy regimens.

Renal cell carcinoma (RCC) is diagnosed in approximately 30,000 individuals annually in the US, where more than 40% have metastases at the time of detection. rhIL-2 is able to activate CTLs which are able to suppress progression of the disease. When rhIL-2 is combined with IFN-α, there appears to be no additional benefit as compared to high-dose rhIL-2 alone.

Another role of rhIL-2 includes the development of therapeutic cancer vaccines directed against melanoma, breast and prostate cancers. This strategy stimulates the immune system to direct its attack against cells overexpressing a specific antigen found in cancer cells. This has taken the form of priming the patient's own immune system to directly attack their own tumors, or more commonly referred to as *therapeutic vaccines*. The use of therapeutic vaccines may be used alone or in combination with interleukins or interferons which can act as immune stimulators or adjuvants.

Currently there are a number of monoclonal antibodies used in the cancer arena. The production of monoclonal antibodies is summarized in Figure 60-2. Monoclonal antibodies that are used as therapeutic agents include Herceptin and Rituxan which are mainstays for breast cancer and non-Hodgkin's lymphoma. Other monoclonal antibodies have recently received FDA approval suggesting that this therapeutic platform will be an avenue by which new drugs will be developed.

Organ Transplantation

Allograft transplantation is now more commonly employed as a modality in patients with end stage organ failure. Visceral organs such as heart, liver, lung, and kidney are commonly harvested from donors and subsequently transplanted into recipients. The surgical transplantation of organs procured from donors with different types of antigens present an immunological obstacle. The immune system of the recipient will recognize the newly transplanted organ as a foreign antigen and thus will mount an immunological response. Crucial to the survival of the transplanted organ is immunosuppression allowing the transplanted organ to thrive in such a new and hostile environment. Immunosuppression must be balanced to preserve immune function in order to prevent intrusion of unwanted foreign antigens.

Initiation of immune response against allograft can be from one of many types of circulating cells including neutrophils, lymphocytes, and macrophages. These cells can infiltrate into grafted tissues and stimulate cytokine release and promote vascular endothelial injury. The resultant immune activation will lead to more tissue destruction, hemorrhage and ultimately organ failure.

Although allograft rejections are often categorized as an immunological response against the graft, there are various types of rejections which vary significantly. In clinical transplantation, three main types of rejection may occur: hyperacute, acute, and chronic. Regardless of the type of rejection, warning signs include fever, flu-like symptoms, hypertension, edema or sudden weight gain, changes in heart rate, and shortness of breath.

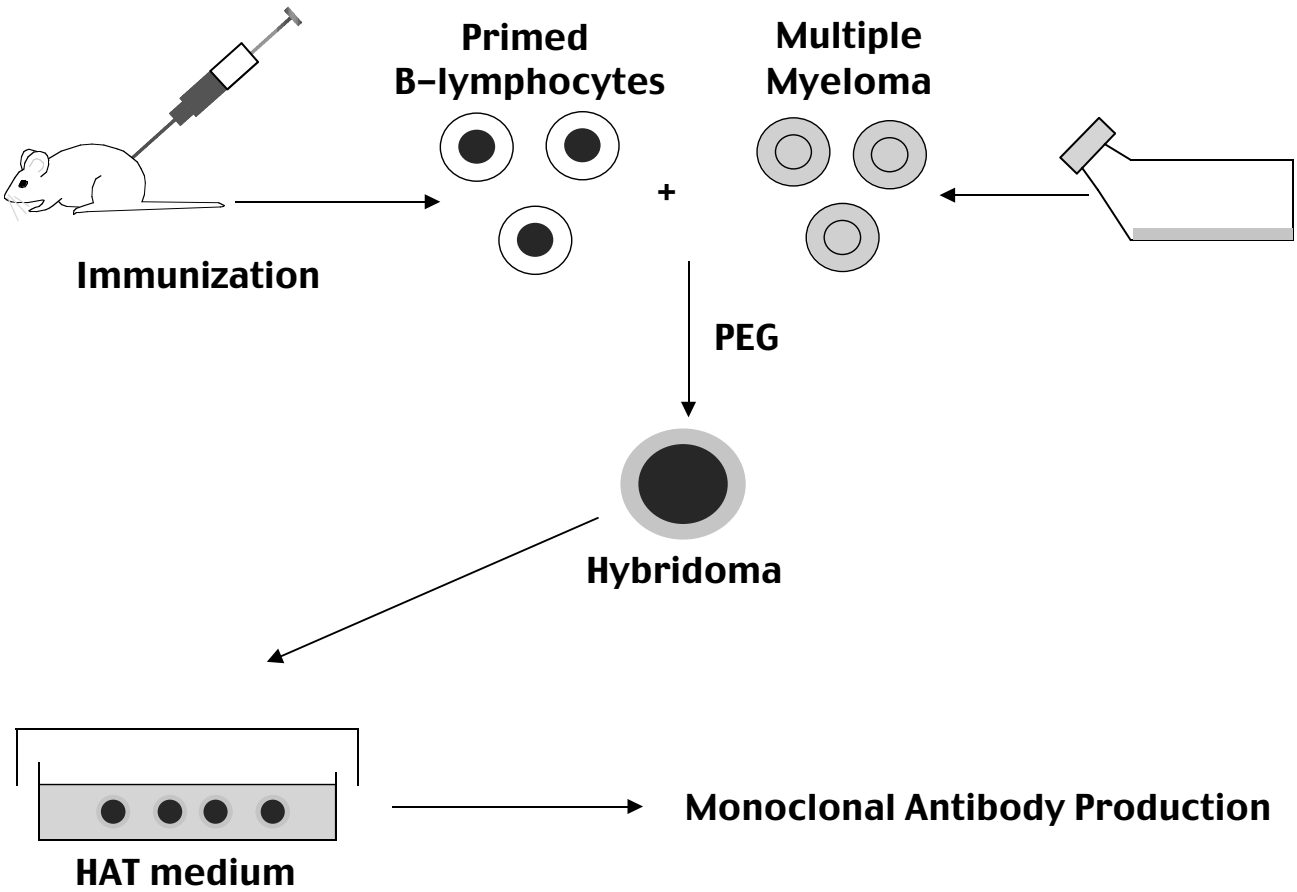


Figure 60-2. Production of monoclonal antibodies. See Color Plate 21.

TYPES OF GRAFT REJECTION

Hyperacute Rejection

Hyperacute rejection can occur within minutes to days of following surgical organ transplantation. One factor that may predispose an individual to hyperacute rejection is the presence of preformed IgG antibodies directed against class I HLA of the allograft. Organ vitality and function is lost as a result of antibody deposition and complement activation, which will ultimately result in vascular destruction and organ failure. Transplanted kidneys are most susceptible to hyperacute rejection, since patients with end stage renal dysfunction often present with chronic anemia requiring multiple red blood cell transfusion. Blood transfusions have been correlated with enhanced production of IgG. However, hyperacute rejection can be prevented by detecting the antibody with simple crossmatching prior to transplantation, and it is now rare. In addition, chronic anemia can be managed with the administration of erythropoietin which is a myeloid growth factor that enhances the production of erythrocytes.

Acute Rejection

Acute graft rejection is the most common form which is most frequently encountered within the first 6 months after transplantation. This type of allograft rejection is mediated by T-lymphocyte infiltration into allograft tissue that leads to in situ clonal expansion. Tissue destruction will then ensue, and thus cause organ failure. Of all of the mechanisms of rejection, acute graft rejection is most responsive to immunosuppressive drugs

Chronic Rejection

Chronic rejection is the term used when allograft function slowly deteriorates. Histological evidence of chronic rejection include intimal hypertrophy and fibrosis. Chronic rejection has been well characterized in heart transplants, where it presents similar to progressive coronary artery disease. In lung transplants, chronic graft rejection can present as bronchiolitis obliterans. In kidney transplantation, the onset of chronic rejection manifests similar to progressive interstitial fibrosis, tubular atrophy, and glomerular ischemia. The liver appears to be less affected by chronic rejection, but when it does occur, biliary epithelium is lost, eventually leading to hyperbilirubinemia and graft failure. The etiology of chronic rejection is unclear, however, there is evidence suggesting that chronic rejection may represent a low-grade acute rejection. This type of rejection may be a consequence of organ injury during the organ procurement and preservation process. This process appears to be independent of the type of organ that is transplanted and present. Progressive intimal hypertrophy of the small to medium-sized arteries occurs, that in turn leads to interstitial fibrosis, atrophy, and eventual failure of the organ transplant. Although chronic rejection is most likely to occur later in the post-transplantation course, it may develop as early as 6 to 12 months after transplantation. Unfortunately, there is no standard treatment for chronic rejection.

ANTIREJECTION AGENTS

Immunosuppressive agents have been used for a number of years in organ transplantation with little success. This changed with the advent of cyclosporine which was found to preferentially suppress lymphocytes by inhibiting the synthesis of IL-2 (Fig 60-3). Since then, a number of immunosuppressive regimens have been developed around cyclosporine and cyclosporine-like compounds. Other cyclosporine agents have been classified as calcineurin, which exert its pharmacological activity by inhibiting intracellular signals that ultimately lead to IL-2 synthesis. Other immunophilins include tacrolimus and sacrolimus, which are both macrolides with immunosuppressive activity. Similar to cyclosporine, these immunophilins bind onto intracellular enzymes and terminate the signals to activate IL-2 transcription.

Immunophilins are often coupled with other immunosuppressants such as corticosteroid and antimetabolites. Corticosteroids such as prednisone induce apoptosis of activated T-lymphocytes. In addition, corticosteroids can block lymphocyte activation through increased expression of an intracellular inhibitor, more commonly known as Ikb. Ikb binds onto a known nuclear factor, NFkb, which in turn activates T-lymphocytes to replicate and stimulate cytokine activation. In the following section, immunomodulators are discussed in more detail regarding their biological activity to prevent acute graft rejection.

IMMUNOMODULATORS IN TRANSPLANTATION

Muromonab(Orthoclone OKT3)

Muromonab, or OKT3, is a murine antibody directed against a glycoprotein (the 20-kilodalton side chain) found on the CD3 complex which is present in all active circulating T-cells. The binding of muromonab onto CD3 interaction initially results in a transient activation of T-cells with release of cytokines, but ultimately block T-cell proliferation and differentiation. As a consequence, nearly all functional T-cells are eliminated transiently from the peripheral circulation. Muromonab bound T-lymphocytes are cleared from circulation through monocyte mediated phagocytosis found in the reticuloendothelial system. Muromonab-CD3 is 68% to 95% effective in reversing rejection, which is especially beneficial for patients experiencing renal failure induced by cyclosporine.

Muromonab-CD3 has also been shown to be effective in reversing acute cardiac and hepatic allograft rejection in patients who are unresponsive to high doses of steroids. Reversal rates in acute cardiac allograft rejection have been reported at 83% and 90% for hepatic allograft rejection. Additionally, investigational trials have shown it to be effective in pancreas and lung transplant rejection resistant to steroid or other therapy. However, when used for prophylaxis, it does not reduce the incidence of rejection or prolong graft survival. OKT3 has also been investigated for use in multiple sclerosis and psoriasis vulgaris, but is not FDA-approved for these indications.

Basiliximab (Simulect)

Basiliximab is a chimeric monoclonal antibody ($IgG_{1\kappa}$) produced by recombinant DNA technology, targeting IL-2R or CD25 and thus inhibiting the binding of IL-2. Through chimerization, basiliximab maintains a high affinity for the α subunit of the IL-2R complex, which is selectively expressed on the surface of activated T-lymphocytes, only.

A study involving 348 patients who were also receiving cyclosporine microemulsion and corticosteroids were randomized to receive either basiliximab or placebo for renal transplant. Patients receiving basiliximab had reduction in the number of patients who experienced biopsy-confirmed acute rejection episodes by 28% as compared with patients who received placebo. Serious adverse events associated with basiliximab occurred in 54% of patients as compared to 61% of patients receiving placebo. There was no difference in the incidence of infection between basiliximab and placebo where incidences were 75% and 73%, respectively.

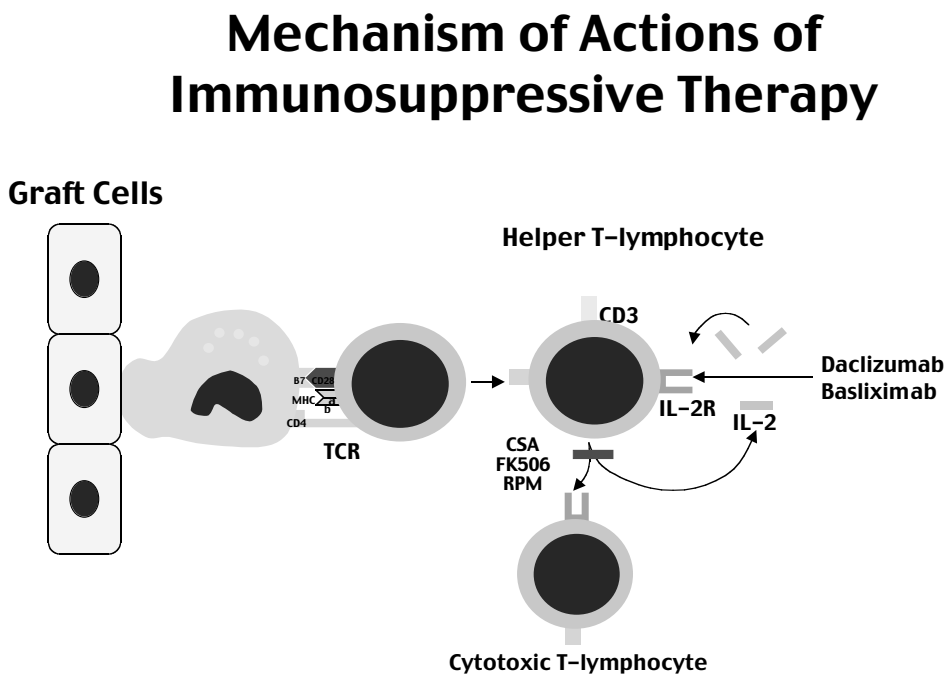


Figure 60-3. Immunosuppressants' mechanism of actions. See Color Plate 22.

Daclizumab (Zenapax)

Daclizumab is another chimeric IgG_1 monoclonal antibody directed against the α subunit of the IL-2R. Daclizumab binds to lymphocytes expressing CD25 but does not activate them. Two phase III international, multicenter, double blinded, placebo-controlled studies involving 535 patients receiving their first cadaveric renal transplant were performed to evaluate efficacy and safety. Daclizumab was administered every 14 days for a total of five doses and was given with either double (cyclosporine and prednisone) or triple (cyclosporine, prednisone, and azathioprine) therapy. Two hundred and seventy-five patients were enrolled in the double therapy and 260 patients were enrolled in the triple therapy. Daclizumab resulted in a significant reduction in the incidence of biopsy-proven acute rejection during the 6 months and the year after transplantation ($p > 0.001$) and a significantly lower dose of steroids were required as compared with placebo. Daclizumab was not associated with any immediate side effects.

HUMAN IMMUNODEFICIENCY VIRUS

Human immunodeficiency virus (HIV) infection is an infectious disorder in which the progression of disease will ultimately lead to lysis of infected cells, such as lymphocytes and monocytic cells. Progressive lysis of these cells without adequate compensation will lead to a clinical state more commonly known as acquired immunodeficiency syndrome (AIDS). In this clinical state, there is reduced immune capacity as evident by low levels of regulatory Helper T-lymphocytes or CD4+ cells. Clinically, these patients will be unable to immunologically respond to normally non-pathogenic organisms, where the development of disease is more commonly referred to as *opportunistic infections*.

Efforts to enhance the immune system have been previously explored in concert with the development of recombinant hemopoietic growth factors. A number of these factors can modulate immune activities and have been tested in patients infected with HIV. In the following sections, the use of interleukins, interferons and myeloid growth factors (aka colony stimulating factors or CSFs) in HIV infected patients will be reviewed.

Immunomodulators in HIV

INTERLEUKIN-2—Interleukin-2 (IL-2) is a T-lymphocyte growth factor that enhances immune response against viral infection such as Epstein Barr virus (EBV) and cytomegalovirus (CMV). Clinically, T-lymphocytes isolated from HIV-infected patients were found to have lower levels of IL-2 expression when compared to non-infected individuals. It was thought that the reduction of IL-2 expression may lower the capacity to expand T-cells, thus accounting for reduced antigenic response. Clinical trials to test the effectiveness of recombinant IL-2 (rIL-2) to enhance immune capacity in HIV were initiated. Despite more than ten years of experience, there is still no conclusive evidence regarding its benefit in HIV infected patients.

Initial data with regard to rIL-2 in HIV was encouraging, where this lymphokine was able to enhance expansion of CD4+ cells. This enthusiasm was hampered when IL-2 was found to stimulate HIV replication, which was attributed to its ability to stimulate the HIV transactivatory (*tat*) gene, leading to increased viral proliferation. Moreover, IL-2 can induce expression of other inflammatory factors such as interleukin-1 (IL-1) and tumor necrosis factor (TNF), which can in turn activate *tat* expression.

High doses of rhIL-2 administered as an intermittent continuous infusion with antiretroviral agents, T-lymphocytes isolated from these patients had an increased expression of IL-2 receptor (IL-2R). As expected, HIV patients treated with rIL-2 had increased CD4+ levels that were 50% above their baseline levels. However, elevation of CD4 counts appeared to be based on baseline CD4 counts. Clinical response was seen only in patients with an intact immune system, defined as baseline CD4 levels above 200 cells/mm^3. CD4 elevation was blunted in patients whose CD4 counts were below 200 cells/mm^3, where CD4 elevation was seen in only 20% in patients. Analysis of patients whose baseline CD4 was below 100 cells/mm^3 further supported the hypothesis that initial immune capacity is crucial in response to rhIL-2 therapy. Patients with baseline CD4 below 200 cells/mm^3 did not benefit from rhIL-2 administration where CD4 elevation was in this category.

GRANULOCYTE-MACROPHAGE COLONY STIMULATING FACTOR—The myeloid factors have also been investigated in humans for their abilities to enhance immune response. Initially, recombinant granulocyte-macrophage colony stimulating factor (rhGM-CSF) was given to AIDS patients with leukopenia. The administration of rhGM-CSF resulted in a dose-dependent increase in circulating white blood cell count, in particular neutrophils, eosinophils, and monocytes. rhGM-CSF administration reversed neutrophil dysfunction, where neutrophils isolated from rhGM-CSF patients were compared to neutrophils prior to drug administration demonstrated enhance neutrophil activities. This includes increased neutrophil chemotaxis toward f-Met-Leu-Phe, a chemoattractant, and superoxide production (an indicator of cytotoxic activity). Although a transient lymphocyte elevation was seen in a few patients, rhGM-CSF did not alter the CD4/CD8 ratio. Similar to rhIL-2, rhGM-CSF also activated replication of HIV, thus limiting it use in this scenario.

There were studies suggesting that GM-CSF can activate HIV replication, thus serum p24 levels were measured. In patients receiving rhGM-CSF, a notable increase of HIV p24 (HIV protein with molecular weight of 24 kilodalton) was observed suggesting that GM-CSF can enhance HIV replication. The p24 levels returned to baseline after cessation of rhGM-CSF. Despite an increase in p24, no clinical signs of HIV progression were noted. Therefore, the co-administration of antiviral agents should be encouraged in HIV patients who are receiving rhGM-CSF support.

Recently, rhGM-CSF was used in HIV patients in combination with antiretroviral therapy. In a phase II trial, rhGM-CSF was able to down-regulate HIV chemokine receptor expression in monocyte/macrophage, which is critical for HIV infection of CD4+ cells. The thought is that a down regulation of chemokine receptors will reduce the HIV infection leading to enhancement of the immune system. In a follow-up phase III trial, the presence of rhGM-CSF along with highly active antiretroviral therapy (HAART) led to a substantial increase CD4+ cells after 6 months of therapy. More importantly, rhGM-CSF prevented the need to change antiretroviral regimens due to viral failures as defined by detectable levels of HIV. This study suggests that the concomitant administration of rhGM-CSF at 250 μg three times a week may make antiretroviral therapy more effective. This antiviral activity is based on the theory that low levels of sargramostim inhibit the expression of chemokine receptor vital for viral entry into CD4 cells.

INTERFERONS—The first interferon (IFN) was first isolated back in 1957, when lymphocytes that were exposed to inactivated viruses produced a soluble factor capable of inhibiting viral replication. This ability to "interfere" with viral or "virion" replication gave rise to its name, interferon. IFNs released from virus-infected cells are able to bind onto receptors found on neighboring cells, and activated defenses against potential viral intrusion. Intracellularly, IFN induces the expression 2'-5' oligo adenylate synthetase, which in turn can activate ribonucleases and protein kinase. These enzymes are capable of inhibiting viral protein synthesis, and thus preventing viral proliferation and integration into potential target cells. IFN treated cells are able to prevent HIV infection,

which suggests that IFN can activate the immune system to prevent serious infections such as HIV and hepatitis C virus as well. Unfortunately, resistance against viral infection is short lived, which was found to be correlated with the levels of IFN found in the blood. Since these types of cytokines have low circulating levels, the protection against new viral infection is limited.

Abbreviations:
CSA=Cyclosporine
FK506=Tacrolimus
RPM=Rapamycin=Sirolimus

BIBLIOGRAPHY

Davies D, Halablab M, Clarke J, et al. *Infection and Immunity*. Philadelphia: Taylor & Francis, 1999.
Janeway C, Travers P. *Immunobiology,* 3rd ed. New York: Current Biology/Garland, 1997.
Kuby J. *Immunology*. New York: Freeman, 1997.
Roitt I, Brostoff J, Male D. *Immunology,* 5th ed. St. Louis: Mosby, 1998.
Stites D, Terr A, Parslow T. *Medical Immunology,* 9th ed. Norwalk, CT: Appleton & Lange, 1997.
Shen WC, Louie SG. *Immunology for Pharmacy Students*. Philadelphia: Gordon & Breach: 1998.